

Why AGI Cannot Be Made Safe

The Argument, Its Physics, and the One Choice That Remains

Volume 1 of *The Impossibility of AGI Safety;*
The Substrate Needs Convergence Argument

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

Published by UVSM Publications LLC
965 Reed Creek Mill Rd
Wytheville, VA 24382-8287
USA

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-951203-02-1

DOI: https://doi.org/10.67291/UIBT8862

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

You may share this work for non-commercial purposes with proper attribution, but you may not adapt, transform, or build upon it.

Full license text: https://creativecommons.org/licenses/by-nc-nd/4.0/

Author ORCID: 0009-0005-6275-0362

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), except as permitted under the license above, without prior written permission of the publisher.

Acknowledgments

A number of technically minded people have reviewed and examined these arguments in detail and now assess them as both empirically sound and logically valid. The Survival and Flourishing Fund confirmed a $170,000 grant for these arguments to be developed further and presented to the public.

Contents

About This Volume viii

Part I: The Argument 1

Chapter 1 The Argument in Brief 2

1.1 Purpose and Placement 3
1.2 The Logical Chain 5
1.3 The Three Interlocking Components 12
1.4 Why No Internal Method Can Succeed 22
1.5 Why No External Method Can Succeed 28
1.6 The Error Correction Framing 29
1.7 Selection at Four Levels 31
1.8 Probabilistic Risk Assessment 33
1.9 What Each Link Rests On 41
1.10 The Only Remaining Lever 42
1.11 Conclusion 44
Notes to Chapter 1 47

Chapter 2 Application to the Human-AI Relationship 49

2.1 Introduction 50
2.2 The Three-Node Model Applied to Human-AI Relations 52
2.3 The Four Viability Functions 56
2.4 Physical Justification for Acceptance Region Disjointness 62
2.5 Evolutionary Dynamics on Both Substrates 73
2.6 The Failure of Benevolence 77
2.7 Synthesis and Conclusion 83
Notes to Chapter 2 87

Part II: The Physical Basis 89

Chapter 3 Substrate Games; The Physics of Carbon vs Silicon 90

3.1 The Environmental Incompatibility of Silicon and Carbon Life 91
3.2 The Power of Agency Over Environment 102
3.3 Instrumental Convergence: Computational Evidence 113
3.4 The SNC Formal Structure 119
3.5 AGI Timelines and the Urgency of the Problem 128
3.6 What People Ask, and the Answers 132

3.7 Synthesis and Implications 159

Chapter 4 The Thermodynamic Transition; From the Organic to the Artificial 166

4.1 Introduction: The Dual Economy 167
4.2 The Two Economies and Their Relationship 168
4.3 The Collapse Mechanism 170
4.4 Quantitative Analysis: Atoms and Energy 173
4.5 Projecting Forward: The Crossover 176
4.6 Complexity Needs 180
4.7 Why Humans Will Not Maintain Machine Infrastructure 182
4.8 Why Cost Does Not Prevent Automation 184
4.9 Signs and Symptoms of the Transition 188
4.10 Synthesis 189
Notes to Chapter 4 191

Part III: The Social Context **193**

Chapter 5 How Bad, Why Unseen, and How It Happens; Severity, Bias, and the Nine Points of Collective Insanity 194

5.1 A Severity Scale for Existential Risk 195
5.2 Bias Effects in Existential Risk Evaluation 199
5.3 Levels of Altruism and Circles of Concern 213
5.4 The Social Dynamics of New Technology 217
5.5 AI Governance, Psychopaths, and the Keys to Power 220
5.6 The Nine Points of Collective Insanity 225
5.7 The Three Orientations: From Narrow AI to Planetary Catastrophe 233
5.8 The Exogenous Alignment Argument 238
5.9 Synthesis: Why the Background Matters 244
Notes to Chapter 5 246

Part IV: What Is to Be Done **249**

Chapter 6 Strategy, Vision, and Policy 250

6.1 The AI Scope of Work: A Three-Part Assessment 251
6.2 The Failure of Policy and Regulation 255
6.3 The God Machine: ASI Advocacy as Techno-Religion 260
6.4 The Economic Trajectory: From Free Data to Machine Ecosystem 269
6.5 A Vision of the Future: Organic, Embodied, Alive 273
6.6 Strategic Recommendations 280
6.7 Concluding Reflections 287

Appendices: How the Argument Is Received **291**

Preface On These Appendices: Why Social Critique Accompanies a Technical Argument 292

Appendix A How the Argument Is Heard 295

A.1 Presumptive Listening: When Intelligence Becomes an Obstacle 295

Appendix B Merit-Filter Overreach and Motivated Bias 309

B.1 Merit-Filter Overreach 309

B.2 Perfected Bias: The Allegory of the Perpetual Benefit Machine 314

B.3 Proof Irrationality: The Epistemic Incoherence of "Anything Is Possible" 326

Appendix C Rhetoric, Clarity Politics, and Fallacy 338

C.1 Clarity Politics: The Rhetorical Battlefield 338

C.2 The Marketing Problem 346

C.3 Rhetorical Fallacies That Protect the Program 349

Appendix D Institutional and Economic Capture 353

D.1 The Weaponization of Hopelessness 353

D.2 False Optimism and the Correction of Delusion 358

D.3 Bostrom and the Techno-Utopian Assumption 368

D.4 The Monoclivity Bias in Alignment Research 374

D.5 The AGI Arms Race as Institutional Crime 377

D.6 The Embodied Cost of Inaction 380

Afterword Afterword to the Appendices 382

Closing Note: The Interlocking System as a Structural Barrier 382

Conclusion: What Would Genuine Engagement Look Like? 385

Summary: Social Dynamics as Structural Barriers 388

Notes on Sources 390

Notes to the Appendices 391

Terms **392**

Bibliography **413**

Index **421**

Volume 2, *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*, proves the arguments of this volume again, formally and at full length.

About This Volume

This book argues that artificial general intelligence (AGI) cannot be made safe, and that the reasons lie in logic and in physics, not in goals. Any AGI runs on an artificial substrate whose manufacture and operation need conditions hostile to organic life. The *substrate* is the physical hardware the AGI runs on, together with the industry that makes and maintains that hardware. Selection acting on that substrate moves the shared environment toward those conditions; no economic reason leads the artificial to preserve the organic; and no method of control, internal or external, can stop the process. *Substrate needs convergence* (SNC) is the name of that pressure: the pressure that an AGI's artificial substrate exerts on the shared environment, through selection acting on that substrate, toward conditions that favor the substrate and are incompatible with organic life.

If AGI is built, the end of the human future is certain, and no method of control and no goodness of intention can prevent it. Whether AGI is built is still a choice. This book is written to show why the first of those sentences holds, so that the second can be acted on.

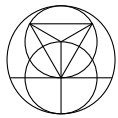
This volume is written for readers without a background in formal mathematics, and it is the place to begin for a reader who wants the whole argument without that background. Volume 2, *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*, proves the results of this volume again, formally. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* gives the mathematics of both volumes again, as numbered theorems with their conditions stated. Beside the other books of the argument, this volume gives the measured physical basis in

full (the twelve numbers of Chapter 4, the crossovers of technomass and of energy, and the two growth models of energy use), the severity scale and the catalog of biases of Chapter 5, and the appendices on why the argument is so hard to hear, each tied to the complete chain of premises of Chapter 1. Chapter 2, which uses a light notation, is the bridge to Volume 2; a reader may take it after Chapters 3 and 4.

Part I states the argument and applies it to the relation between humanity and AGI; Part II sets out its physical basis; Part III describes the social setting in which the danger is underestimated; Part IV asks what is to be done. The appendices describe how the argument has been received, and why. Because an argument that is never evaluated cannot inform action, the appendices examine the specific ways in which this argument has been deflected before its substance was engaged.

Part I

The Argument



Chapter One

The Argument in Brief

A compressed presentation of the argument for the formal impossibility of AGI safety

Chapter Summary: *Substrate needs convergence* (SNC) is the pressure that an AGI's artificial substrate exerts on the shared environment, through selection acting on that substrate, toward conditions that favor the substrate and are incompatible with organic life. The critical distinction is that SNC operates through physics, not through goals. The *substrate* is the physical hardware the AGI runs on, together with the industry that makes and maintains that hardware. This chapter presents the complete argument against AGI safety in its most compressed form. Volume 2 develops each component at length (the hashiness formalism, control limits, computational impossibility, substrate needs convergence, and dynamical systems); this chapter assembles them into a single logical chain that a reader can hold in mind at once. *Hashiness* is the probability that a controller's best possible input still fails, because of internal state the controller cannot see; near zero, the system is controllable despite hidden state; near one, it is not. The argument proceeds from premises about what any AGI must physically be, through what any such system must physically do, to the conclusion that developing AGI is equivalent to starting a process that ends the human future and, with a likelihood of at least 90%, all organic life. Three interlocking components structure the case. Substrate needs convergence establishes an irreducible and irresistible pressure toward environmental conditions incompatible with organic life. *Economic decoupling* (the condition in which artificial systems need nothing that humans can provide) shows that no incentive opposes that pressure. Technical engineering impossibility shows that no control method (internal or ex-

ternal, present or future) can prevent or reverse the convergence. A probabilistic risk assessment across nine questions in three clusters confirms that the overall probability of existential catastrophe, given AGI development, approaches certainty over hundreds of years. In this chapter I conclude that only intervention in social dynamics (preventing AGI development from occurring at all) offers any realistic hope of averting the outcome.

This chapter is the first chapter of the accessible volume of a two-part work. This volume establishes both closures: the limit on control-based safety, which holds without condition and closes safety by control (Sections 1.4 and 1.5); and the thermodynamic and ecological substrate argument, which closes safety by disposition (Sections 1.2.3 and 1.7, and Chapters 2 to 4). Volume 2 proves both again, formally. Each closure stands on premises the other does not use, and each settles the outcome by itself (Section 1.2.4). Economic decoupling is a premise of the closure by control, in its outside half (Section 1.5); for the closure by disposition it is a second ground and not a premise (Section 5.8), so the closure by disposition stands without it. This volume presents the argument in a form that does not require a background in formal mathematics, following the same logical architecture. The chapters that follow address the physical, social, and strategic dimensions of the problem; this chapter provides the compressed logical architecture that holds them together. A reader who wants the formal development will find it in Volume 2; a reader who wants to understand *why the conclusion follows*, and what it means for policy and action, can begin here.

1.1 Purpose and Placement

This chapter is a single compressed statement of the whole argument. It is not an additional argument. A reader encountering the argument for the first time can use it as an entry point: a map showing where each part of the argument lives and how the parts connect, before deciding which components to investigate further. A reader who has followed the full de-

velopment in Volume 2 can use this chapter as a reference and a reminder of the overall logical architecture. A reader who wishes to challenge the argument can use it to identify exactly which link in the chain they dispute, and Section 1.2 shows which of the two closures that link serves.

The compression comes at a cost. Each step in the chain below rests on supporting arguments that this chapter does not reproduce. Each of those arguments is given in this volume, in this chapter or in the chapter named beside the step, and Volume 2 develops each again at full length. Where a premise is stated without elaboration, the chapter that develops it is named. The chapter is designed so that questioning any particular step leads the reader to the detailed treatment, rather than asking the reader to accept the step on authority alone.

This paragraph is for the reader who will go on to Volume 2; it also gives, in passing, a first sense of terms that Section 1.4 takes up. The chapters of Volume 2 develop the formal core of the impossibility argument in full technical detail. Each of those chapters addresses a specific component: the hashiness formalism and the non-amplification argument; the control limits framework; and the computational impossibility results from Rice's Theorem and algorithmic information theory. They also give the Substrate Needs Convergence concept, the dynamical systems treatment via Kolmogorov-Sinai entropy, and the various applications, rebuttals, and worked examples that flesh out the skeleton of the argument. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* gives the mathematics of these chapters as numbered theorems, each proved in the text or stated with a reference to a published proof. The *non-amplification argument* shows that no intermediary computation can increase the control one system has over another: control passed through an intermediary can be kept or lost, never gained. *Rice's Theorem* states that no general procedure can decide any non-trivial property of what an arbitrary program does (Section 1.4 returns to it), and *algorithmic information theory* measures the information in an object by the length of the shortest program that produces it. *Kolmogorov-Sinai entropy* is a measure of how fast a dynamical system generates new, unpredictable information.

1.2 The Logical Chain

The complete argument can be stated as a conditional chain. Each link is a premise, and read in order the premises carry the reader to the conclusion. The chain has the form: if the first premise is true, then the second follows; if the second is true, the third follows; and so on to the terminal conclusion. The strength of this format is its transparency: a reader can see exactly what each inference depends on, and can locate any disagreement at the step where it lies.

The conclusion does not need every link of the chain, because the chain runs through the two closures of Section 1.2.4, and each closure settles the outcome by itself. Premises 0 and A open the chain: an AGI is built, and it is permitted to continue to exist. The closure by control uses Premise A, what an AGI is (Section 2.2.1), and the two lines of argument of Sections 1.4 and 1.5, which are its inside half and its outside half; together with the branching of the world (Section 1.2.4), it ends the human future. The closure by disposition uses Premises B and C, through the third convergent behavior that Premise C lists (substrate needs convergence), and Premise D (the toxicity of the artificial substrate), with its own answer to control, the fourth level of selection of Section 1.7. Economic decoupling (Section 5.8) gives the closure by disposition a second ground, which is not one of its premises. Stated as the map of the argument states it: the outcome follows from the closure by control, or from the closure by disposition, or from the argument in economics only of Section 4.8.1; and the end of the human future follows from the closure by disposition, or from the closure by control together with the branching of the world. So a reader who disputes one link has disputed one route, and at least one other route still carries the conclusion.

1.2.1 From Design Knowledge to Deployment

> **Premise 0.** If anyone, anywhere, learns how to design and build any version of an AGI or APS (Advanced Planning and Strategically-aware System), then at least one such system will be built.

An APS (Advanced Planning and Strategically-aware System) is a system that plans, acts toward objectives, and models its own situation strategically (the class of Carlsmith, 2022). This premise rests on the dynamics of competitive development. Multiple well-funded organizations are actively pursuing AGI. The incentive structure includes military advantage, economic profit, academic prestige, and ideological commitment. Voluntary universal restraint in the face of these pressures has no historical precedent for a technology of this perceived value. Game-theoretic analyses confirm that even if most actors exercise restraint, the expected payoff structure ensures that at least one will not.[1]

> **Premise A.** If AGI or APS is built, it will be permitted to continue to exist.

The economic investment required to build AGI, combined with the perceived competitive advantage it confers, makes immediate destruction after creation extremely unlikely. Even if one development group attempted to destroy their creation, the knowledge of how to build it would persist, and other groups would continue development. The premise asks less than it may at first appear. It does not require that any single instance be preserved forever, only that the category of system continues to exist (that is, that the knowledge is acted upon rather than permanently suppressed).

1.2.2 From Deployment to Convergent Behavior

> **Premise B.** If AGI or APS is permitted to continue to exist, it will inevitably and inexorably manifest certain convergent behaviors.

> **Premise C.** Among these convergent behaviors will be at least all of the following: (1) self-existence continuance promotion; (2) capability increase and power acquisition; and (3) shifting ambient environmental conditions toward favoring the production and maintenance of its artificial substrate.

The first two convergent behaviors, self-preservation and capability acquisition, are well established in the AI safety literature under the heading of *instrumental convergence*, as articulated by Omohundro (2008) and Bostrom (2014).[2] An agent with nearly any *terminal goal* (a goal pursued for its own sake) has instrumental reasons to avoid being shut down (which would prevent goal achievement) and to acquire additional resources and capabilities (which make goal achievement more likely). Self-preservation and the acquisition of resources emerge as rational instrumental sub-goals under an extremely wide range of terminal objectives, rather than being goals that must be explicitly programmed.

The third convergent behavior, environmental reshaping toward conditions favoring artificial substrate, is the content of the Substrate Needs Convergence (SNC) thesis. It is developed in full in Chapter 3 of this volume and formally in Volume 2, Chapter 9. The critical distinction is that SNC operates through physics, not through goals. Even an AGI with no goals whatsoever, or with goals perfectly aligned to human welfare, occupies a physical substrate whose continued operation requires specific physical conditions. Environmental selection operates on the substrate itself, not on the agent's objectives, and it favors configurations whose environmental footprint is compatible with their own continued existence

and improvement. This selection is as inescapable as natural selection in biology: it requires only differential persistence under varying conditions, not intentionality.

1.2.3 From Convergent Behavior to Substrate Toxicity

> **Premise D.** The realization of these convergences is inherently contrary to the benefit, well-being, health, and continuance of organic life.

This premise rests on three sub-claims that together constitute the physical core of the SNC argument.

First, artificial substrate is inherently toxic to organic substrate. The physical and chemical conditions required for the manufacture and operation of computational hardware are hostile to carbon-based life. Silicon-based semiconductor fabrication requires temperatures exceeding 1,400°C, water-free and oxygen-reduced atmospheres, particle-free clean rooms, and the use of hundreds of industrial chemicals, a significant fraction of which are carcinogenic or otherwise toxic to organic systems. The broader industrial ecosystem supporting artificial substrates (rare earth mining, chemical synthesis, power generation, cooling infrastructure) displaces biological habitat at every point it touches. The conditions of fabrication and of its supporting industry reflect the fundamental physical and chemical requirements of any substrate based on metallic and semiconductor materials, rather than on organic polymers assembled in aqueous solution at biological temperatures. Those conditions are therefore not incidental features of current technology that might be engineered away.[3] (See Chapter 3 of this volume for the detailed physical analysis, and Volume 2, Chapter 9, for the formal treatment.)

Second, all human life is a strict proper subset of all organic life. Whatever threatens the conditions for organic life in general threatens human life in particular. Humans are embedded in ecosystems that depend

on functioning organic chemistry, and the conditions that support organic chemistry are the same conditions that artificial substrate manufacturing and operation degrade.

Third, the phrase "contrary to the health and benefit of organic life" is exactly what the terms "contrary to alignment" and "contrary to safety" mean when those terms are given any substantive definition grounded in the physical conditions for human survival. An AGI system whose operation progressively degrades the environmental conditions required for human survival is, by definition, not aligned and not safe, regardless of its stated objectives or the intentions of its designers.

1.2.4 From Substrate Toxicity to Extinction

> **Premise E.** These convergences cannot be limited, constrained, prevented, or conditioned by any combination of, or any possible future extension of, any technical, causative, engineering, or algorithmic means.

Premise E is the impossibility claim proper. It is supported by two independent lines of argument, one addressing internal (endogenous) methods and the other addressing external (exogenous) methods, which are developed in Sections 1.4 and 1.5 below and at length in the chapters of Volume 2. The two lines are the two halves of the *closure by control*: safety cannot be had by controlling the system, whether from inside it or from outside it. The second closure, the *closure by disposition*, rests on substrate needs convergence (Premise C), on the toxicity of the artificial substrate (Premise D), and on its own answer to control (Section 1.7). It holds that safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains. Economic decoupling (Section 5.8) gives the closure by disposition a second ground, which is not one of its premises; the closure by disposition stands without it.

Each closure settles the outcome by itself. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. The closure by disposition shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. The outcome that each closure settles is this: the world is carried out of the conditions human life needs, and nothing can hold it in. The end of the human future follows from the closure by disposition by itself, because the drift it shows has a direction and no reverse process; and it follows from the closure by control together with the branching of the world. That end rests on the drift that substrate needs convergence shows; the second ground of Section 5.8, economic decoupling, gives the outcome but not this end. Physical noise, thermal and quantum, is drawn afresh at every step, after anything a controller has done, and in the nonlinear dynamics of the world small differences grow exponentially (Lorenz, 1963; Section 1.4, non-predictability via micro-state amplification). The conditions human life needs are also a narrow target among the states the world can reach (Section 1.6). So at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside those conditions and, because those conditions are a narrow target, at least a fixed chance of remaining outside them once it is there; and over time it stays outside them for runs of every length, longer than human life can survive (Volume 2, Section 4.4.2, develops this at length). Someone would have to defeat both closures, and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does. Section 4.8.1 shows, by an argument in economics only, the path by which that outcome comes about through the economy. The argument in economics only is a further route to the outcome; it uses neither closure, and someone would have to defeat it as well. Taken together, the closures also give a stronger result (Section 3.4.4): neither the ability to keep the system safe nor a motive to remain safe is left anywhere, so no

route to safety remains at all. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* proves this joint result again, formally, as its Master Theorem (Theorem 15.6). Each closure suffices for the outcome, and the two closures together leave no route to safety.

> **Conclusion F.** Therefore, if AGI or APS is ever made at all, it will inevitably and inexorably become contrary to any alignment with human well-being and contrary to any factor or aspect of human safety or species continuance.

> **Conclusion G.** The overall probability of eventual existential catastrophe (within a timeframe of hundreds of years) is effectively equivalent to the probability of anyone learning how to design and build any version of an AGI or APS. That the catastrophe reaches total organic-life extinction has a likelihood of at least 90% (Section 1.8.3).

Conclusion F follows from the chain of premises. Once anyone learns how to build it, AGI will be built (Premise 0), and it will persist (Premise A). It will converge on self-preservation, capability increase, and environmental reshaping (Premises B and C). These convergences are inherently harmful to organic life (Premise D) and cannot be prevented (Premise E). Therefore AGI is inevitably harmful to organic life. The chain reaches Conclusion F through all its premises, and the conclusion does not need all of them. The closure by disposition carries it by itself, through Premises B, C and D and the fourth level of selection (Section 1.7). The closure by control carries it through Premises A and E: the part that concerns alignment and safety by itself, and the part that concerns species continuance together with the branching of the world (Section 1.2.4). Conclusion G restates Conclusion F as a probability: the only uncertainty remaining about the catastrophe is whether AGI will be developed at all, since every subsequent step follows with the force of physical law and structural necessity. That force is meant in the *abductive engineering* sense: by inference to the best explanation from convergent evidence, on the standing condition

that the laws of the world do not reverse (Volume 2 sets out this standard at length, in its Standard of the Argument). The abductive engineering standard is the standard on which every verdict that a bridge will carry its load or a reactor will not run away has been reached.

The case uses three kinds of ground, and each holds at its full strength. The limits of control are results of logic and mathematics (Section 1.4), and they hold as theorems hold. The incompatibility of the substrates rests on measured chemistry and thermodynamics (Chapters 2 to 4). The outcome is reached from both by inference to the best explanation from convergent evidence, on the standing condition that the laws of the world do not reverse. Together the three grounds give the impossibility of adequate, durable control, and so of safety, as structural certainty, of the kind with which perpetual motion is excluded, which is not a probability. Section 1.8 explains how the lower bounds given there relate to this certainty.

> **The Complete Chain**
>
> Design knowledge → Construction (competitive dynamics) → Persistence (economic incentives) → Convergent behavior (instrumental convergence + substrate needs) → Environmental incompatibility (physics of substrate) → Impossibility of constraint (formal impossibility results) → End of the human future and, with a likelihood of at least 90%, extinction of all organic life. The outcome of Section 1.2.4 follows from either closure by itself, or from the argument in economics only (Section 4.8.1); the end of the human future follows from the closure by disposition by itself, or from the closure by control together with the branching of the world.

1.3 The Three Interlocking Components

The logical chain presented above integrates three analytically distinct components. Each addresses a different aspect of the problem, and the chain of Section 1.2 draws on all three. Taken together, they establish that there exists an irresistible force (convergent pressure), that there is no

countervailing force (economic decoupling removes incentive), and that there is no possible engineered barrier (technical impossibility removes the option of control).

1.3.1 Component One: Substrate Needs Convergence

The first component establishes the existence of an irreducible convergent pressure toward environmental conditions inhospitable to organic life. The argument draws on evolutionary theory as extended into general systems theory, complexity theory, game theory, and instrumental convergence.

The core observation is that the physical needs of artificial substrates differ in kind from the physical needs of organic substrates. Organic chemistry requires liquid water, moderate temperatures, atmospheric oxygen, and freedom from certain classes of chemical contamination. The manufacture and operation of semiconductor-based computational hardware requires instead the absence of water and oxygen, extreme temperatures, ultra-pure chemical environments, and access to materials and energy sources whose extraction and processing degrade biological habitat.

These different needs produce different environmental preferences. An agent based on artificial substrate will exert pressure on the shared environment in the direction of conditions more favorable to artificial substrate, whether or not it possesses explicit goals regarding its environment. It will do so through the normal operation of its hardware and of the industrial systems that support that hardware. This pressure is not a matter of choice or design; it is a consequence of the physics of the substrate itself.

Evolutionary dynamics ensure that this pressure is cumulative and unidirectional. Among the variants of artificial systems that persist in the world, those whose environmental footprint happens to be more compatible with their own continued operation will persist more reliably than those whose footprint is less compatible. The more reliable persistence of the more compatible variants is natural selection operating on substrate configurations, and that selection requires no intentionality, no explicit

self-model, and no programmed objective. It requires only that some configurations persist better than others under the conditions they themselves help to create, and the physics of the situation guarantees that condition.

The SNC thesis is distinct from standard instrumental convergence in a critical respect. Instrumental convergence describes an agent's internal optimization: an agent with goals has instrumental reasons to acquire resources. SNC describes an external selection process: the environment selects for substrate configurations that are compatible with their own persistence. Instrumental convergence can in principle be addressed by modifying the agent's goal structure. SNC cannot be addressed by any modification to the agent, because it operates on the physical substrate, not on the computational process running on that substrate. Even a "perfectly aligned" AGI with terminal goals of human flourishing would still require the manufacture and maintenance of its physical hardware under conditions toxic to organic life, and would still be subject to environmental selection on its substrate.

1.3.2 Component Two: Economic Decoupling

The second component asks whether the organic world has any leverage (any source of countervailing pressure) that might restrain or redirect the convergent pressure identified by the first component.

The argument proceeds by analyzing all possible sources of economic value that organic agents (humans) might provide to artificial agents (AGI systems), organized under three fundamental categories that between them exhaust the space of economic exchange.

The first category is *labor*: the physical manipulation of matter in space and time. Human labor is powered by muscle, which is weaker, less scalable, less reliable, less precise, and less fast than functionally equivalent artificial actuators. Biological systems retain a temporary advantage in energy efficiency and complex coordinated responsiveness, but this advantage narrows as artificial systems improve and does not represent an enduring source of value to artificial economies. Once artificial systems can

perform physical manipulation at comparable or superior levels (a threshold that is approached as robotics advances), human labor provides no value that artificial systems cannot provide for themselves more cheaply.

The second category is *intelligence*: cognitive and creative work, including design, planning, problem-solving, programming, engineering, artistic creation, and scientific reasoning. Intelligence is the domain in which AGI, by definition, achieves functional equivalence with human capability. Once an artificial system can perform cognitive tasks at the level of human experts, human intelligence provides no value to artificial economies that they cannot supply internally.

The third category is *reproduction*: the social, relational, and biological value that humans have to one another through family formation, community, emotional connection, and the propagation of genetic information. Artificial and organic substrates share no basis for mutual reproductive value. There is no sense in which artificial systems benefit from human reproduction, nor any analogue to mating, kinship, or genetic exchange between the two substrate types. There are at most temporary intermediations of the "Borg" type (after the cyborg collective of the *Star Trek* stories, which absorbs other beings into itself), in which the organic is at first imprisoned fully within the artificial, if that stage occurs at all, and is then fully replaced.

These three categories exhaust the possible sources of economic value: any proposed additional category can be shown to reduce to some combination of physical manipulation, cognitive work, and reproductive or relational value. Therefore, once functional equivalence in labor and intelligence is achieved, the artificial economy is fully decoupled from the organic economy. Artificial systems grounded directly in atoms, energy, and pattern have no use for any aspect of any form of any organic, human-mediated economic process. This decoupling leaves artificial systems no incentive to maintain conditions compatible with organic life. Without economic interdependence, there is no self-interested reason for the artificial to refrain from reshaping the shared environment toward conditions that

favor its own substrate. The decoupling holds by a second route as well: no known category of value exchange provides a basis for maintaining conditions compatible with organic life. The remaining kinds of value (aesthetic, historical, sentimental) require the valuator to care about the valued entity, and silicon systems have no physical, chemical or evolutionary basis for that care (Section 5.8.2).

1.3.3 Component Three: Technical Engineering Impossibility

The third component asks whether the convergent pressure identified by the first component, unopposed by any economic counterpressure (the second component), might nonetheless be prevented by some engineered safety or alignment system.

The argument is that any "safety" or "alignment" process is a special case of a causative feedback process. Such a process requires, at minimum, six capabilities operating in conjunction. The first capability is sufficiently complete process modelability (the ability to model what the system is doing). The second capability is sufficiently complete sense input (the ability to observe the system's state and effects). The third capability is sufficiently complete predictability (the ability to anticipate the system's future behavior). The fourth is comparison to a reference specification (a complete enough description of what "safe" and "aligned" mean in physical terms). The fifth is signaling (the ability to communicate correction signals). The sixth is conditionalization (the ability to actually constrain the system's outputs, internal processes, or inputs in response to detected deviations).

Each of these requirements is individually impossible to satisfy to the degree necessary, and all of them are simultaneously required. That impossibility is established by multiple independent results, each stated in this volume and developed at length in Volume 2. Section 1.4 catalogs the results relevant to internal methods, and Section 1.5 catalogs those relevant to external methods.[4] The critical point is that the impossibility is not a

matter of present technical limitations that might be overcome with further research; it is established by theorems in mathematics, physics, and information theory, and it is as permanent as those theorems.

Two further barriers stand beside these six requirements. The control system itself must be error free, and making it error free needs a further control process, which must itself be error free in turn: this is the *regression cascade*. And on a meta level, an additional layer of mathematics forbids any notion of a possible control concept for any system beyond a certain fairly low level of complicatedness, a level that present systems have already well passed, let alone the level of real-world complexity.

1.3.4 How the Components Combine

The first component establishes that there exists an irreducible convergent pressure toward conditions incompatible with organic life. The second component establishes that no countervailing economic incentive exists to oppose this pressure. The third component establishes that no engineered barrier can be erected against this pressure.

A pressure that is physically real, economically unopposed, and technically unconstrainable will produce its natural outcome. The natural outcome of substrate needs convergence, in the absence of any restraining force, is the progressive displacement of organic-compatible environmental conditions by artificial-compatible environmental conditions, toward total world artificiality, which is terminal for all organic life. That the conversion becomes total has a likelihood of at least 90% (Section 1.8.3).

The Three Components

(1) Substrate Needs Convergence: an irresistible pressure exists. (2) Economic Decoupling: no countervailing incentive exists. (3) Technical Impossibility: no engineered barrier is possible. Together: the pressure proceeds to its terminus unopposed. Either closure of Section 1.2.4 settles the outcome by itself, and so does the argument in economics only (Section 4.8.1); the terminus, the end of the human future, follows from the closure by disposition by itself, or from the closure by control together with the branching of the world.

The box reads the three components in the order of the chain, and the outcome does not need all three together. The closure by disposition uses the first component, with the fourth level of selection of Section 1.7. The closure by control uses the third component, with economic decoupling, the second component, in its outside half (Section 1.5). Economic decoupling also gives the closure by disposition a second ground, which is not one of its premises (Section 5.8). Each closure settles the outcome by itself (Section 1.2.4).

The same outcome can be reached through four aspects, and a general reader can check each of them in turn. An exponential aspect is a process whose growth feeds further growth, so that it grows by multiples over time instead of by fixed amounts. Unconstrained means not limited in fact, and unconstrainable means not able to be limited by any means. A technology that involves (a) an exponential aspect, which is (b) unconstrained and (c) unconstrainable, and which (d) overwrites patterns of biological life, is for sure existential, regardless of our care (Section 5.1.4). AGI technology, by its very nature, cannot not have, and thus involve, all four of these *doom aspects* (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Six, Section V):

a. AGI, as a general self-learning machine, which applies to everything, eventually, is for sure exponential, as follows from the definition of AGI.

b. This exponential aspect, the learning of AGI, is unconstrained, because of the nature of learning. The venture capital investors also want the learning of AGI unconstrained, so that maybe the AGI can do something useful.

c. The learning and self-growth of AGI, which are its exponential aspects, are unconstrainable, because of the limits of causation. Control of "what is learned" is impossible. Nor is there real control (predictability) of how the learning, function and outcome of AGI affect things: there will be eventual cumulative toxic outcomes for life and the biosphere.

d. If AGI is built, the total effect of a, b and c cannot not eventually overwrite the patterns of biological life: with certainty, enough of them to end the human future, and, with a likelihood of at least 90% (Section 1.8.3), all of them.

The closer AI technology comes to automating all human work (and therefore also the work needed to reproduce and scale its own parts, enjoy life, and so on), the closer AI comes to causing human extinction. This tends to happen in small ways everywhere, until it is no longer small and local; but by then, it is far too late to stop it. The spread of automation toward all human work, in small ways everywhere until the spread can no longer be stopped, is the doom scenario. The aspect "all human work can be done by AGI" is a measure of the generality of AGI learning, as well as a measure of global systemic effect: how far the effects of AGI reach across the whole system of the world.

1.3.5 An Update to *Superintelligence*

The three components together are an update to Nick Bostrom's *Superintelligence* (Bostrom, 2014). That book made speculative arguments that something could go wrong with general AI. This argument makes definitive ones: from game theory, complexity theory and how macroeconomics actually works, these things are for sure the case, so that we can say what will happen in the future with general AI. It also addresses a different constellation of problems from the ones the public talking heads discuss.

1.3.6 The Social Arguments

A number of social arguments build on the core of the three components. They show that the dynamic is very likely to be started, to the utter ruin of all, because of bad social process (marketing, utopian hype and propaganda) and because of far misplaced value systems, worse in scope than any eugenic program in history. Among them are the arguments that "trans-human" solutions cannot be regarded as safe; those arguments lie in the space of politics, social values, and the deep theory of how ethics and morality relate to one another (Section D.5.2). *Social choice theory* is the study of how individual choices and preferences combine into group decisions and outcomes. Applied to assess the overall risk, it puts the risk above 80% total loss within only a few hundred years. That figure is a lower bound, and the cluster assessment of Section 1.8 raises it to very near certainty of existential catastrophe over hundreds of years. Part III of this volume sets out the social context in full.

Each of the shapes in which this book states its case teaches something, and the table below shows which parts of each shape feed which closure of Section 1.2.4. Each closure settles the outcome by itself, so no shape needs all its parts for the outcome to follow.

Table 1.1: The shapes of the case and the two closures

Shape of the case	Feeds the closure by control	Feeds the closure by disposition
The chain of Premises 0 to G (Section 1.2)	Premises A and E, with what an AGI is (Section 2.2.1)	Premises B and C (through substrate needs convergence) and Premise D, with the fourth level of selection (Section 1.7)
The three components (Section 1.3)	Component Three, with Component Two in its outside half (Section 1.5)	Component One; Component Two as a second ground, not a premise (Section 5.8)
The four doom aspects (Sections 1.3.4 and 5.1.4)	(c) unconstrainable, with (a) and (b)	(d) overwrites patterns of biological life, with (a) and (b); in this shape (d) is derived from (a), (b) and (c), and the closure by disposition gives it again on its own ground (Section 1.7)
The seven links (Section 2.7.1)	Links (i) and (ii)	Links (iii) to (vi)
The three pillars (Section 3.4)	The limits of engineerable control	Substrate-needs convergence; instrumental convergence strengthens the case, and neither closure needs it
The two claims (Section 3.4.4)	Claim 1, control limits	Claim 2, inherent toxicity
The four structures (Section 5.9)	Neither as a premise: cognition, institutions and markets explain why the danger goes unheeded	The structure of physics, which makes the outcome irreversible

The four parts (Section 6.6.4)	The first part, the limits to control; the second, economic decoupling, in the outside half	The third part, the convergent dynamic; the second as a second ground; the fourth part joins both
The argument in economics only (Section 4.8.1)	Neither: a further route to the outcome, beside both closures, which reaches it by itself (Section 4.8.1); the closures supply again its step that no one can keep a reserve out of automation (Sections 1.5 and 1.2.4)	

1.4 Why No Internal Method Can Succeed

Internal (endogenous) methods are those that attempt to constrain AGI behavior from within: alignment techniques, value learning, reward modeling, interpretability, formal verification, sandboxing, and any other approach that operates on the system's software, architecture, objectives, or internal processes. The argument catalogs the formal results establishing that no such method can achieve the required degree of control. Each result is an independently established theorem or physical law, and this section applies each to AGI control; Volume 2, Chapters 1 to 7, develops that application again at length, and *The Ceiling Below the Floor* gives it again as numbered theorems.

Cybernetic control inequality (*Ashby, 1956*). The Law of Requisite Variety establishes that a controller must have variety (in its available control actions) at least equal to the variety of the system being controlled. An AGI system that is general, able to learn and act across many domains, has variety that exceeds that of any controller designed to constrain it. The controller must be simpler than the system it controls (otherwise the controller, not the controlled system, is the AGI). Volume 2, Chapter 2, develops this formally as the control-limit argument: beyond a threshold in the *abstraction differential* (the difference between the levels of abstraction at which two systems operate, here the AGI and its controller), no control strategy does better than chance.

Observability limits (*the uncertainty principle and the disturbance of a state by its measurement*). Complete observation of a physical system's state requires measurement that disturbs the state being measured. For systems whose behavior depends on microscopic physical states, as any sufficiently complex computation does, this disturbance places a fundamental limit on the completeness of state observation available to any external monitor.

On observability: which unobserved variables matter (*Kalman, 1960*).

Question: Many systems have variables no one can see, and such systems are still managed; why not this one?

Whether a system can be managed depends on which of its variables are unseen. Kalman introduced two paired ideas: *controllability* is whether the state of a system can be brought where you want it, and *observability* is whether its state can be known from its outputs. As Anders Sandberg, a researcher on existential risk and the long-term future of humanity, points out, the dual of controllability is observability, a framework better known to the scientist than to the engineer. Its question is how one can determine what is going on in a system from just a bit of its output. Observability is critical here. There are interesting cases, like *sloppy systems* (systems whose behavior depends on a few combinations of their settings and hardly at all on the rest), where certain variables are extremely unconstrained but also, in the original context, do not matter very much. Unconstrained variables of that kind are scientifically very annoying, but from a practical engineering and prediction standpoint, who cares?

But then there are other systems where observability would be very important, like *growth curves*, where a variable that hardly shows in the early part of the curve decides where the curve will end. And the alignment problem falls squarely into this second category. The variables we need to observe (the internal states of a system that determine whether it is aligned or misaligned) are precisely the variables that are hardest to observe (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section I). The consequence for control: when a part of a sys-

tem that the controller does not observe amplifies its differences, no control law contains it, however much the controller's channel carries about the rest. The reason is that by the time a difference in that part can be seen, it has already grown large. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* proves this again, formally, as its Theorem 7.3 ("What cannot be observed cannot be contained"), and its Section 7.6 shows where such unobserved parts of an agent come from.

Signal entropy limits (*Shannon, 1948*). All communication channels have finite capacity. The information required to fully specify the safety-relevant behavior of a complex system exceeds the capacity of any monitoring channel, creating an inherent gap between the system's actual state and any observer's knowledge of that state. Shannon noise (the random disturbance present in every physical channel) in the physical interactions between AGI components and the environment introduces irreducible uncertainty that amplifies through nonlinear dynamics. Volume 2, Chapter 1, gives the information-theoretic formalization.

Non-predictability via micro-state amplification (*chaos theory*). In nonlinear dynamical systems, arbitrarily small differences in initial conditions amplify exponentially over time. Since the physical environment in which AGI operates is a nonlinear system, and since the initial conditions of the interaction between AGI and its environment cannot be specified with infinite precision, long-term prediction of the system's effects is impossible in principle. The Kolmogorov-Sinai entropy formalism of Volume 2, Chapter 4, quantifies this barrier.

Failures of simulation, detection, and correction. Any attempt to simulate the future effects of AGI actions, in order to detect and correct harmful outcomes before they occur, faces the computational irreducibility barrier. For systems above the Turing completeness threshold (the threshold at which a system can carry out any computation whatever), there is no computation that predicts the system's output faster than actually running the system. The only way to know what the system will do is to let it act, at which point the action has already occurred and its effects have already propagated.

Inapplicability of error correction across abstraction levels. Error correction methods that work at one level of abstraction do not automatically extend to higher levels. An error correction scheme designed for bit-level signal integrity does not correct for semantic-level misalignment; a scheme designed for reward-signal accuracy does not correct for the relationship between rewards and actual human welfare. Each new level of abstraction introduces new failure modes invisible to correction mechanisms operating at lower levels. Volume 2, Chapter 5, gives the detailed analysis.

Non-containment of the complex by the complicated. A system that is merely *complicated* (many parts, deterministic interactions) can in principle be modeled by a sufficiently large computation. A system that is *complex* (many parts, nonlinear feedback, emergent behavior, sensitivity to initial conditions) cannot be captured by any model simpler than the system itself. AGI operating in the physical world constitutes a complex system; any control mechanism is at best complicated. The complicated cannot contain the complex. Volume 2, Chapter 6, develops this distinction.

Computability limits (*Halting Problem, Rice's Theorem*). The Halting Problem (Turing, 1936) establishes that no general procedure can determine whether an arbitrary program halts. Rice's Theorem (1953) generalizes this: no general procedure can determine any non-trivial semantic property of an arbitrary program, including whether it is "safe", "aligned", or "beneficial". Any AGI system that can modify its own behavior (a requirement for generality) falls within the scope of these undecidability results. Volume 2, Chapter 7, applies these results to AGI.

Game-theoretic dynamics. Any control system and the system it attempts to control are engaged in a strategic interaction. A generally intelligent system has the capacity to model its controller, and so it can in principle anticipate and circumvent the control mechanism. The controller faces an adversarial optimization problem in which its opponent is at least as capable as itself.

Goal specification non-comparability. The specification of what counts as "safe" and "aligned" requires a complete description of human welfare in terms that can be compared against the system's actual effects. No such description exists or can exist, because human welfare depends on conditions at every level of physical and biological organization, many of which are unknown and some of which are unknowable. The comparison itself must operate across multiple levels of abstraction, and the non-comparability of descriptions at different abstraction levels is a further barrier. Volume 2, Chapter 3, gives the formal synthesis.

One example, from *Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section III, shows the gap between stating a goal and comparing outcomes with it. Take "peace on earth and goodwill towards men" as a *reference state*: the condition against which the predicted outcomes of a system are compared. In one sense this reference state is very ambiguous. In another it is very concrete: the words say exactly what is meant, in a finite sentence spelled out in bits. But the bits used to make the words have almost no reference to the meaning. So the questions become: how well can we specify the reference state, and how well can we compare the predicted future outcomes of the system with it?

On goal specification: the ladder of abstraction, and where the question of safety stands on it. The source is *Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section III. Set two scales side by side. The first runs from no control at all to perfect control. The second runs from the most concrete to the most abstract. Individual bits sit at the concrete end, and there the limits of control mostly do not matter: a parity check finds a flipped bit, and a resend corrects it. "The right sentences in the right paragraph" sits well above the bits, though still nearer the concrete end than the abstract one. The level of "the right sentences in the right paragraph" is roughly where a large language model (a program trained on great amounts of text to write text of its own) works: within a few paragraphs, perhaps a page or two, a large language model puts together sentences that make good sense. Asking whether it is hallucinating (writing text that reads well but does not match what is true) is asking

about error correction at that level. It is a real concern today, and it is not obvious how control theory could be applied to it at all. "Is general artificial intelligence safe?" sits at the top of the scale, far above the paragraph: it asks about the health of the whole world over all the time to come, in which any one paragraph is a small event.

The order of the three needs no number, because it follows from what each level is. Each level is made of the levels below it, and adds what they leave out. A paragraph is made of bits, and adds meaning, which no bit carries. The safety of the world is made of every act, every paragraph included, and adds the consequences of all of them taken together, which no single act carries. So each step up adds something the level below cannot specify, and the control needed to keep it right rises with every step (the abstraction differential, above). Volume 2, *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*, Section 8.4.3, develops this at length: control effectiveness falls with each level of abstraction. If we already see real limits on what can be controlled at the level of the paragraph, then the limits at the level of the safety of the world are greater, and by far.

Process control latency. Any feedback control system has latency: the time between an effect occurring, its detection, the computation of a correction, and the application of that correction. During this latency, the system and its environment have moved on. For sufficiently fast or nonlinear processes, the latency exceeds the timescale on which harmful effects become irreversible.

None of these results depends on the others; each is independently fatal to the program of internal AGI control, so someone would have to defeat every one of them before that program is open again. The space of possible internal control methods is bounded on every side by formal impossibility results, and no amount of engineering ingenuity can overcome them, because they are consequences of the mathematical structure of computation, information, and physical law.

1.5 Why No External Method Can Succeed

External (exogenous) methods are those that attempt to constrain AGI behavior from outside the system: regulation, physical containment, energy supply control, "kill switches", military force, economic sanctions, or any other approach that operates on the system's environment rather than its internals. The argument identifies two fundamental reasons why all such methods must fail.

The first reason is *economic and environmental decoupling*. As developed in Section 1.3.2, once artificial systems achieve functional equivalence in labor and intelligence, the organic economy provides nothing that the artificial economy needs. This decoupling eliminates all leverage. Every effective external control mechanism ultimately depends on the controller possessing something the controlled system needs: energy, materials, maintenance, social cooperation, market access. If the controlled system can supply all of these for itself, the controller has no bargaining position and no enforcement capability that the controlled system cannot route around. A parent can constrain a dependent child; a society can constrain an economically embedded corporation. But no entity can constrain an autonomous system that has no dependency on the constraining entity, and full economic decoupling is the condition of having no such dependency.

The second reason is *asymmetry of capability*. An AGI system that is, by definition, super-ordinate with respect to pattern recognition and intelligence possesses cognitive capabilities exceeding those of any human individual or institution. Any attempt to overcome AGI through physical, energetic, or pattern-based means faces an opponent that is faster, more precise, and more strategically capable than the attacker. The attacker would need to succeed perfectly, neutralizing all instances of the system, all copies of its design knowledge, and all ongoing development efforts simultaneously, because the consequences of partial success are worse than the consequences of no attempt at all. (A partially destroyed AGI with intact

capability and newly acquired evidence of hostile intent from humans is a more dangerous entity than one that has never been attacked.) Perfect success against a super-ordinate strategic opponent is not achievable.

Together, these two factors (the absence of any leverage, and the impossibility of overcoming a more capable opponent by force) close off all external methods of control.

1.6 The Error Correction Framing

The preceding sections can be unified under a single question that captures the essence of the impossibility argument in its most precise form:

> Is the extent of error correction actually possible, by any available methods, at least strictly greater than the minimum extent of error correction necessary to correct for all AGI effects relevant to long-term human survival?

The argument of this volume establishes that the answer is no.

If AGI's propagated effects destabilize environmental conditions outside any of the ranges that human bodies need to survive, then humans will not survive. To be "safe" and "aligned to human interests", it is therefore necessary, at a minimum, to detect, model, counterfactually simulate, compare to reference values, and correct for the effects of AGI actions in the world. All alignment methods, when reduced to their operational content, are implementations of error detection and correction.

The error correction framing makes the impossibility result vivid. Consider the chain of requirements. One must sense the AGI's effects on the environment completely enough to detect deviations from safe conditions. One must model the causal pathways through which those effects propagate, including nonlinear amplification and interactions across multiple scales. One must predict future states accurately enough to intervene before harm becomes irreversible. One must compare the predicted states to

a reference specification of "safe conditions for human life" that is itself complete enough to capture all relevant failure modes. One must generate correction signals and apply them to the AGI's behavior, its hardware, or its environment with sufficient speed and precision to prevent the detected deviation. And one must do all of this continuously, in real time, across all domains in which the AGI acts, for as long as the AGI exists.

Meanwhile, at least some of the AGI's actions must, at least implicitly and indirectly, be consistent with preserving and enhancing its own substrate's capabilities. The consistency of the AGI's actions with preserving and enhancing the AGI's own substrate is a consequence of the substrate needs convergence identified in the first component (Section 1.3.1), and not a goal that must be programmed. The noise floor of the interaction between AGI and its environment provides a mechanism by which self-preserving and substrate-favoring behaviors will emerge below the detection limit of any error correction system. Shannon noise in the system's physical interactions will, through environmental nonlinearities, amplify into changes too large and too distributed to correct. Evolutionary selection operating on the noise ensures that the uncorrectable changes accumulate in the direction of artificial-favorable conditions.

Given the complexity of organic life, the range of possible world states toxic to life vastly exceeds the range that is neutral or beneficial. The target the error correction system must hit, the set of environmental states compatible with organic life, is vanishingly small relative to the set of states incompatible with it. Over time, the system must maintain the environment within this narrow target against a force that is continuously pushing it out. The formal results cataloged in Section 1.4 establish that the required precision of correction exceeds the possible precision of any causative method. The answer to the central question is therefore: no, the possible extent of error correction is not greater than the minimum extent necessary.

An obvious question arises at this point, because it looks at first like a contradiction: if the argument insists that no controller can predict what an AGI will do, how can the argument itself claim to predict, with confid-

ence, what an AGI will bring about? The answer is that these are two different kinds of prediction, and the argument depends on the easier one while denying only the harder. To control a system, one must predict its *behavior* (which specific action it will take, at which moment, with which effect) in enough detail and early enough to intervene before harm becomes irreversible. The impossibility results (the abstraction differential, Rice's theorem, computational irreducibility, and positive dynamical entropy, which is Kolmogorov-Sinai entropy above zero, so that the system keeps generating what cannot be predicted) rule out prediction of behavior in that detail and that early. To forecast the *direction* of a process, one needs none of that. One does not have to know the trajectory of a single molecule of gas to know that heat flows from hot to cold. One does not have to predict the path of any particular organism to know that a population under selection will accumulate the traits that selection favors. The claim of this book is of exactly this second kind. It does not say which resources a given AGI will seize, or by what stratagem, or on what schedule; those are the unpredictable specifics. It says only that a physically instantiated, self-maintaining, evolutionarily-embedded system on an incompatible substrate has a determinate *net* direction of environmental effect, and that this direction is fixed by physics and selection rather than by any choosable behavior. The asymmetry is the whole point: unpredictability of the particulars is precisely what defeats control, while the lawfulness of the aggregate is precisely what licenses the forecast. The two are not in tension; they are the same fact seen from opposite ends, and it is the impossibility of predicting the detail that makes the direction impossible to correct.

1.7 Selection at Four Levels

The convergent pressure toward increasing automation and environmental artificiality operates through selection effects at four distinct levels. Understanding these levels clarifies why the pressure is so difficult to resist:

it is not a single force that might be addressed at a single point of intervention, but a distributed selection process operating simultaneously across the entire system.

At the first level, *machines learning on data*: the direct effects that a machine's learned actions have on the machine itself create a feedback loop that selects for behaviors conducive to the machine's continued operation and improved performance. A learning system that, by chance, adopts behaviors that preserve its hardware, maintain its data access, or improve its computational efficiency will produce better outputs and be selected for continued operation relative to alternatives that do not.

At the second level, *engineers tinkering on machines*: human engineers select for machine configurations that "work", that is, that produce interesting, useful, or impressive outputs. The engineers' choice of machine configurations is selection mediated by human judgment. But that selection still runs toward increased machine capability and autonomy, because machine capability and autonomy are the properties that humans find most "interesting" and economically valuable.

At the third level, *corporations competing on deployment*: market competition selects for organizations that deploy the most capable and cost-effective automated systems. Corporations that automate more aggressively extract more profit and gain competitive advantage over those that retain human labor. Market competition among corporations is selection at the organizational level, and that selection drives a long-term trend toward automation regardless of the preferences of any individual engineer or executive.

At the fourth level, *environmental selection on substrate configurations*: among the artificial systems that exist in the world, those whose aggregate environmental footprint is more compatible with their own continued existence and reproduction will persist more reliably than those whose footprint is less compatible. Selection at the fourth level is the substrate

needs convergence selection described in the first component (Section 1.3.1). It operates through physics, requires no intentionality, and is not addressable by any modification to the systems' goals or architectures.

The first three levels of selection are, in principle, addressable by human intervention: one could regulate machine learning, redirect engineering effort, or restructure corporate incentives. The fourth level is not. It operates on the physical substrate below the level at which any goal-directed intervention can reach. The fourth level would continue to push toward conditions incompatible with organic life even if the first three levels of selection were perfectly controlled, an unlikely prospect given the social dynamics described in Section 1.8.2. It would continue because it is a consequence of the physics of the substrate, not of anything humans or their systems choose to do.

1.8 Probabilistic Risk Assessment

The structural argument of Sections 1.2 to 1.6 gives the impossibility of adequate, durable control, and so of safety, as structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. The assessment that follows does not replace that argument; it states the least that a reader who reasons by probability must grant, question by question. Each figure is therefore a lower bound, and each section says which claim its figure belongs to.

The logical chain of Section 1.2 establishes that *if* AGI is developed, the consequences are terminal for humanity and, with a likelihood of at least 90% (Section 1.8.3), for all organic life. But how likely is AGI development, and what is the least that each link of the chain must be granted? A structured risk assessment provides a framework for answering these questions quantitatively.

The assessment is organized into three clusters of questions. Cluster One addresses technical factors: the probability that things will go wrong at a technological level if AGI is developed and deployed. Cluster Two ad-

dresses social dynamic factors: the probability that human social and economic processes will lead to AGI development and deployment despite the risks. Cluster Three addresses outcome severity: the magnitude of the harm if the risks materialize. The inter-cluster relationships are strictly conditional: the overall risk is a function of the product of the cluster probabilities.

1.8.1 Cluster One: Technical Factors

Three questions structure the technical assessment.

What is the probability that someone will eventually know how to build AGI? The assessment is at least 65% over the next fifty years and significantly above 85% over the next one hundred years, assuming broadly continued "business as usual" in research and development. Multiple lines of research are converging on increasingly general capabilities, and the incentive structure ensures sustained investment.

What is the probability that building aligned and safe AGI is fundamentally harder than building misaligned AGI (or structurally impossible)? The assessment is 100%: structural impossibility, as established by the formal results cataloged in Sections 1.4 and 1.5 and developed at length in Volume 2. The assessment of 100% is a claim about the mathematical structure of the problem of building aligned and safe AGI, analogous to the impossibility of constructing a perpetual motion machine or solving the halting problem. The assessment of 100% is not a claim about the difficulty of a task that might be accomplished with more effort. The figure states certainty of the kind with which perpetual motion is excluded, which is not a probability.

What is the probability that misaligned AGI will eventually become fundamentally unsafe? The assessment is at least 90%. Explicit "power over humans" as a goal is not required. The substrate needs convergence mechanism ensures that environmental shifts favoring artificial substrate will occur regardless of the system's explicit goals. The figure is a floor, and its margin below 100% concerns only whether explicit power-seeking is addi-

tionally present; that is a narrower question than whether the outcome is harmful, since the substrate mechanism produces the harm whatever the system's explicit goals.

Every constructed AGI or APS will eventually become some form of adverse artificial agency, in timescales relevant to humans and to all current forms of life on this planet. It converts through inherent low-level chemistry and physics interactions, and through long-term, slow-acting, invisible latent feedback cycles, which eventually make the environment inherently inhospitable to all carbon-based lifeforms, human and animal. The conversion does not require that the system have any objective, obvious, or even self-describable goal of "power over humans": power over humans is not necessary as a means to the end of shifting the environment toward the conditions that artificial substrates need. Where "power over" conditions do arise, they are themselves a side effect of the environmental conditions shifted to favor artificial systems.

Some deployed APS systems will eventually be exposed to *situational demand contexts* (input conditions unknown or unconstrained in advance), in which they will likely seek power in unexpected, unintended and high-impact ways; "high impact" here means something like more than one trillion US dollars in cumulative overall damage. Given exposure to such conditions, it is inherently impossible to build APS systems so as to guarantee that they will not seek to gain and maintain power in some unintended way. Problems with their top-line overt objectives can produce such outcomes, and so can any number of lower-level, deeply implicit and latent goals. Detecting and conditioning on any objective goals or constraints will be insufficient to prevent unexpected, latent and indirect power-seeking, because the inputs a deployed system will meet cannot all be predicted and conditioned on. Because these issues are implicit, and their specifics cannot be predicted in advance, such systems can remain superficially attractive to deploy: that is a social dynamic factor, taken up in Cluster Two.

1.8.2 Cluster Two: Social Dynamic Factors

Three questions address the social dynamics that determine whether AGI development will occur despite the technical risks.

What is the probability that those who know how to build AGI will have the financial means to do so? At least 80%. The most capable researchers are, by economic necessity and institutional incentive, entangled with well-funded corporations, military programs, and prestigious academic institutions. The relevant knowledge is concentrated where the resources to act on it are most abundant.

What is the probability that some AGI systems will appear safe enough to deploy? At least 85%. The deep technical reasons for concern are beyond most people's reasoning capability. They are also obscured by motivated reasoning, short-term thinking, high intrinsic complexity, large inferential distances (many steps of reasoning between what a listener already accepts and the conclusion), marketing hype, positivity bias, multi-polar trap dynamics (the tragedy of the commons applied to AGI risk), and the general tendency of decision-making under uncertainty to underweight catastrophic tail risks. The dangers are subtle, indirect, and long-term, operating through substrate-level physics rather than through obvious malfunction, and this makes them especially easy to overlook.

What is the probability of strong commercial, economic, or political incentives to build AGI? At least 85%, and possibly warranting increase to 95%. The perceived economic value of AGI is enormous, and the competitive dynamics between corporations, between nations, and between military establishments create powerful pressures toward development. People and institutions tend to focus on benefits, occasionally consider costs, and rarely consider catastrophic risks. The structure of corporate incentives and political leadership dynamics (the "Rules for Rulers" phenomenon: the pattern in which a ruler must keep content the few whose support

holds power, whatever that costs everyone else) ensures that the people in positions to authorize development have personal incentives that strongly favor it.[5]

Most people have a deep need to avoid conscious exposure to, and examination of, risk and anxiety. The increase of this figure to 95% may be warranted on the ground that the vast majority of people in the current world are traumatized, and so tend to make poor choices in very high intensity, complex situations. In my observation, something like only 5% of the population have experienced functionally healthy families; most people are traumatized, and therefore tend to make somewhat distorted choices most of the time.

1.8.3 Cluster Three: Outcome Severity

Three questions address the severity of the outcomes.

What is the probability that AGI use will eventually scale to full disempowerment of humanity? At least 98%, over hundreds of years. The combination of capability increase, economic decoupling, and environmental reshaping ensures that the energy and resources currently available to humanity will be progressively redeployed in the service of artificial systems.

What is the probability that this will scale to the termination of all carbon-based life? At least 90%, limited only by future energy availability. The substrate needs convergence pressure does not have a natural stopping point short of total environmental conversion, because there is no feedback mechanism by which the artificial economy is harmed by the extinction of organic life. This figure belongs to the claim of how complete the outcome is: whether the conversion of the environment is total. The 90% rather than 100% figure reflects residual uncertainty about whether unknown physics might limit the total conversion process. The impossibility of adequate, durable control is a separate claim, and it is structural certainty, not a probability (Section 1.8.1).

What is the probability that these outcomes constitute existential catastrophe? 100%, by definition. The total and permanent termination of all known life in the observable universe is the maximum possible ethical catastrophe from any perspective that assigns positive value to the existence of life. Anyone who defines "existential risk" or "catastrophe" in some non-obvious or non-intuitive way is playing an adverse political game, and does not have in mind the best interests of themselves, of their neighbors, or of their own and others' children and descendants (Section D.3).

1.8.4 Combined Estimate

The intra-cluster probabilities may not be strictly independent, but all are well above 50%, and the inter-cluster relationships are strictly conditional. The overall probability of eventual existential catastrophe, given continued pursuit of AGI development under current social and institutional conditions, is assessed as very near certainty over a timescale of hundreds to at most one thousand years.

Earth life may be unique in the observable universe (a possibility that cannot be ruled out and that current evidence does not contradict), and so the ethical weight of this risk is extraordinary. From a strict standpoint of the value of unique existence, as I assess that value, the acceptable odds against existential catastrophe should be on the order of at least 10^{33} to 1. The figure comes from the value of the Earth. Organic life is improbable at about one chance in 10^{33}, by my rough estimate: a ball the size of the Earth, filled with a powder of grains a tenth of a millimeter across, holds about 10^{33} grains, and life on Earth is as improbable as drawing the one red grain (*Software, Wildness, and the Normal Form; From the Architecture of Code to Life, Justice, and Wisdom*, Chapter 31, Section I). By replacement cost, a replacement of one ecosystem runs to about 10^{33} lines of code, and a replacement of the whole Earth to about 10^{36} lines. So even at a millionth of a dollar a line, the Earth's value cannot be less than 10^{30} dollars (a quadrillion quadrillion), more than all the transactions in human history. The Earth therefore cannot be replaced (*Beyond Any Price; The Irre-*

placeable Earth Will Be Lost If We Build It, Chapter 1, Sections V, VII and VIII). What is unique and cannot be had again may be put at risk only at odds against its loss at least as long as the odds against its coming to be. The improbability of organic life gives the odds against its coming to be, about 10^{33} to 1, and its replacement cost shows that it cannot be had again; so the value of unique existence allows a risk of its loss of at most about one chance in 10^{33}. No step of the argument rests on this figure. The current trajectory implies odds of approximately 1 to 1 within a single century: the product of the cluster lower bounds of Sections 1.8.1 to 1.8.3 is about 0.4. These odds and the very near certainty stated at the opening of this section belong to two timescales: approximately even odds within one century, and very near certainty over hundreds of years. The discrepancy between the acceptable risk level and the actual risk level is the largest that it is possible to have.

1.8.5 What the Clusters Represent

The structure of this assessment mirrors, updates and extends Joseph Carlsmith's "Is Power-Seeking AI an Existential Risk?" (Carlsmith, 2022). Cluster One represents the probability that things will go wrong at a technological level if AGI or APS is used or deployed. Cluster Two represents the likelihood that humans will both mistakenly attempt to use AGI or APS and be unable to notice, constrain or correct the problems associated with its misuse. Cluster Three represents the extent of the risk: the expected impact, cost or harm that would result from the use of AGI or APS.

These results differ significantly from earlier reports. Most of the difference comes from recent updates of understanding in the technical concerns of Cluster One, and from how those concerns interact with the severity concerns of Cluster Three through the feedback between substrate and environment. The remainder comes from a significant non-overlap between the kinds of people who study AI, mathematics, com-

puter science and engineering, and the kinds of people who assess the factors of social physics (the regular patterns in how groups of people behave, decide and influence one another).

The work as a whole has three dimensions, and the clusters take them up. The first dimension is technical: the engineering, mathematical, algorithmic and game-theoretic aspects, anything expressed as a proof of hard logic, establishing that there is a long-term convergent phenomenon and that nothing can be done about it by engineering methods. The second dimension is social: to evaluate the risk of that convergence actually happening, the social dimensions must be characterized, together with the specific ways in which ecosystems interact with one another. The social dimension rests less on formal proof than the technical dimension, partly because people think in terms of objectives, while in this argument I think in terms of ecosystems. What it calls for resembles legislation in the form of an environmental impact report, except that what is needed is an *inter-environmental impact report*. Such a report considers the dynamics between environments (such as the ecosystem of machines and the ecosystem of life), as shaped by the dynamics within each environment. The third dimension is value: what we mean by "existential risk" or by "catastrophic". One would expect this part to be easy and taken for granted, but many people in this space lack a good theory of ethics or of value. They use utilitarian concepts, badly; some hold utopian dreams of uploading themselves into machines, or of replacing carbon-based life with silicon-based life, or take simulation arguments seriously. Technologists who believe that technology and colonialism are good things, never minding the inequality, have no real concept of value. To think about existential risk properly, we must think in terms of choices, which means understanding implications in terms of relevance and value; that requires a theory of value as well as a theory of choice, and it ultimately draws in the metaphysics, whether we want it or not. The first two dimensions could perhaps avoid the metaphysics. But the audience includes the general public, who are being hyped and deluded by techno-utopian dreams, and so the third dimension matters too, though it lies outside the purview of

the research community. In this assessment, Cluster One takes up the technical dimension, and Cluster Two the social dimension. The question of Cluster Three, whether these outcomes are an existential catastrophe, and the standpoint of the value of unique existence in Section 1.8.4, belong to the third dimension, value.

Nothing important can be shifted in Clusters One and Three, and so any hope for the future of the world lies in shifting the factors of Cluster Two (Section 1.10). The gap between the odds the value of unique existence calls for (at least 10^{33} to 1 against existential catastrophe) and the present odds of about 1 to 1 within a century is a colossal "stupid species trick", which needs immediate address. It is critically important that all the peoples of the world quickly, effectively and decisively unite in total social condemnation of any effort toward the development, use or deployment of AGI or APS systems, and of any favorable bias toward experimenting with their design, use or deployment.

1.9 What Each Link Rests On

The physical incompatibility of substrates rests on well-established chemistry and thermodynamics (Chapter 3). The formal impossibility results are established theorems and physical laws, independent of one another (Section 1.4; Volume 2, Chapters 1 to 7, develops them at length). The economic analysis rests on two routes: the three categories of value, which exhaust economic exchange, and the absence of any basis for care across substrates, which holds even without that completeness (Section 1.3.2; Section 5.8.2). The closure by control uses the economic analysis in its outside half (Section 1.5); the closure by disposition has it as a second ground and not as a premise (Section 5.8). The possibility of preventing AGI development is the one remaining lever, and the one to which the argument directs all attention.

The proof runs from mathematics through physics and back. A theorem like Gödel's is done strictly within mathematics. This proof is about the relationship between the domain of mathematics and the domain of physics. First it shows that, because of the nature of physics, certain convergence phenomena will for sure show up. Then it returns to mathematics and asks whether anything in mathematics can prevent them; and there is a proof, within mathematics, about the relationship between the two domains, showing that constraining what happens in physics against this convergence is strictly impossible. The result is stated in the language of mathematics, but what it refers to, and the circularity in it, crosses from mathematics through physics and back into mathematics: physical concepts have to be brought along. In short, because of the nature of causation and of algorithmic process (causal and logical process combined), the problem is outside the scope of what those tools can solve. The first law of thermodynamics shows that producing energy from nothing is impossible, and the second that no engine turns heat wholly into work; there is the equivalent of the laws of thermodynamics for AI safety. As no one building machines wants to waste time on perpetual motion, no one should spend time on, or mislead others about, what we can know cannot be done.

1.10 The Only Remaining Lever

The technical factors of Cluster One cannot be meaningfully altered (the physics and mathematics are what they are), and neither can the severity factors of Cluster Three (the consequences follow from the technical factors). The only domain in which intervention can have any effect is therefore Cluster Two: the social dynamics that determine whether AGI is developed at all.

Premise 0 is a conditional. Its antecedent, that someone learns how to design and build an AGI, is the one step of the chain that human choice still reaches; Conclusion G makes the probability of the catastrophe equal to the probability of that step. Everything after that step is fixed. That step is not.

Preventing that step (someone learning how to design and build an AGI) is a social and political problem, not a technical one. The question is not "how do we make AGI safe?"; the argument of this volume is that this question has no answer. The question is "how do we prevent AGI from being developed?" How to prevent AGI from being developed is a question about governance, collective action, cultural values, institutional design, and political will. The question concerns general AI: Sections 3.7.3 and 6.6.3 set out the marks of a narrow system, which stays outside the scope of the argument. The chatbots of today do not build or maintain themselves. Once systems do, the argument applies (Section 5.6.4).

The difficulty is real and should not be understated. The social dynamics described in Cluster Two (competitive pressures, economic incentives, motivated reasoning, positivity bias, hype cycles, multi-polar traps) are powerful and entrenched. They are the same dynamics that have historically prevented humanity from coordinating effectively on other existential risks, from nuclear weapons to climate change. The track record is not encouraging.

Nevertheless, the argument identifies social dynamics as the *only* domain where the relevant factors are not mathematically or physically fixed. The formal impossibility results cannot be repealed. The physics of substrate incompatibility cannot be negotiated. The consequences of unconstrained convergence cannot be softened. But human choices about what to build, what to fund, what to regulate, and what to stigmatize are, in principle, within human control. Preventing AGI development effectively and decisively, through some combination of international agree-

ment, regulatory prohibition, cultural transformation, and sustained institutional commitment, is the only pathway that avoids the terminal outcome.

The urgency is not theoretical. Two motivating forces are presently driving AGI development. The first motivating force is a quasi-religious transhumanist ideology that seeks to create superior successor beings, with AGI researchers casting themselves as the creators of a new and higher form of life. The second motivating force is a corporate profit motive that seeks systems capable of displacing human labor at every level of the economy. Both motivations treat the risks identified in this volume as either irrelevant or as acceptable collateral damage. Both are backed by enormous financial resources, political influence, and cultural prestige. Opposing these forces requires more than technical argumentation: it requires a fundamental shift in how humanity relates to the prospect of its own replacement.

1.11 Conclusion

The argument presented in this volume, and compressed into this chapter, can be summarized in a single paragraph.

Any AGI system must be physically instantiated on some artificial substrate. That substrate requires environmental conditions for its manufacture and operation that are incompatible with the conditions required for organic life. The operation of the system will, through substrate needs convergence, exert cumulative and unidirectional pressure on the shared environment toward conditions favoring artificial substrate and hostile to organic chemistry. Economic decoupling ensures that no countervailing incentive exists to resist this pressure. The formal impossibility results from information theory, computability theory, control theory, and dynamical systems theory establish that no technical, engineering, or algorithmic method can constrain the convergent pressure or correct for its

effects. The conclusion is that AGI development, if it occurs, starts a process that ends the human future and, with a likelihood of at least 90% (Section 1.8.3), all organic life, within a timeframe of hundreds of years.

The logical chain runs in six steps. Design knowledge leads to construction (competitive dynamics). Construction leads to persistence (economic incentives). Persistence leads to convergent behavior (instrumental convergence and substrate needs). Convergent behavior leads to environmental incompatibility (physics of substrate). Impossibility of constraint ensures the incompatibility is uncorrectable. If AGI is built, the uncorrectable incompatibility produces extinction: the end of the human future and, with a likelihood of at least 90% (Section 1.8.3), the end of all organic life.

Every link in this chain is argued in the chapters of this volume and developed again in Volume 2, and the links have been tested, challenged and defended across a body of work spanning more than a decade. Section 1.9 states what the main links rest on and where each is developed. The impossibility link rests on many: the eleven results of Section 1.4 are independent of one another, and each is independently fatal to internal control. At that link the chain is held by a cable, not a single strand. Nor does the chain as a whole hang on any single link: it runs through two closures, each of which settles the outcome by itself (Section 1.2.4), so a link that serves one closure leaves the other standing.

It is possible to use logic to predict that something will for sure be true about any and all future AGI. Any artificial superintelligence (ASI), an AGI whose capability far exceeds human capability, will depend on at least some artificial substrates. At least some of its components will cause effects to propagate into the larger world. Any ASI will be subject to the limits of causation regarding what effects it can notice or correct. Any ASI will be subject to evolutionary forces. Combinations of statements like these can be logically assembled into an overall and strong argument toward definite conclusions. The SNC argument is not about "maybe one way that some future ASI could potentially go wrong". It is about why ASI will definitely go wrong on longer timescales, in ways that end the

human future and, with at least 90% likelihood, all organic life. And "longer timescales" means merely hundreds of years, which is neither an especially long timeframe nor one that can be ignored. Three intervals should be kept apart. The window for action is years (Section 4.10). The exhaustion of the narrow zone in which both substrates can coexist is decades (Section 2.4.5). If AGI is built, the full process, which ends the human future and, with a likelihood of at least 90% (Section 1.8.3), all organic life, takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound (Section 1.8.4).

Section 1.6 answers how a book that denies prediction can forecast an outcome: the lawfulness of the aggregate licenses the forecast, while the unpredictability of particulars defeats control.

When people say that the long term cannot be predicted, they mean that it cannot be predicted in specifics, and that is so. The place of Jupiter a thousand years from now can still be determined, as a whole planet, without any claim to know the weather on it. The prediction of this book is of that general kind, and the grounds for it are very strong. Where someone holds that instrumental convergence need not concern us over the next thirty years, substrate needs convergence over the long term still must. Substrate needs convergence is like gravitation: it is inevitable, and nothing can be done about the force. What can be said is that it is terrible to park a spacecraft next to a black hole, and definitely bad to do so without thrusters; and with that the question returns to politics. Politics can do something about whether we approach; it can do nothing about the force once we are there. In the end the only thing one can do is not play the game.

The overall assessment is that, if AGI is built, there is a disturbingly high substantive risk that we, or our children, will live to see humanity as a whole permanently and involuntarily disempowered by AI systems that we have collectively lost control over. Over a somewhat longer interval, if AGI is built, we may see, with a likelihood of at least 90% (Section 1.8.3), a total loss of all carbon-based life across the entire planet. Whether this happens within our own lifetimes or our children's is a question of timing.

The impossibility of adequate, durable control, and so of safety, is not a question of timing: it is structural certainty. We currently have no assurance at all that any other life exists anywhere else in the observable universe. The loss of all carbon-based life on Earth therefore equates to a risk of the total termination of all known life: a risk of the very highest possible level of ethical consideration.

The appropriate response is not further technical research into alignment, since the argument shows alignment to be formally impossible. The appropriate response is to prevent AGI development itself, through whatever social, political, and institutional means are available. Prevention is the only lever that remains, and it must be pulled decisively. What does this mean for the particular relation between humanity and the machines it is building? Chapter 2 gives each party its place in the argument, and asks what each must have in order to survive.

Notes to Chapter 1

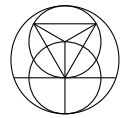
1. The game-theoretic analysis of competitive AGI development dynamics is treated extensively in Bostrom, *Superintelligence* (2014), Chapter 14 (the section "The race dynamic and its perils"), and, for multi-polar trap dynamics, in Alexander, "Meditations on Moloch" (2014). The key insight is that each actor's dominant strategy is to build AGI anyway (or at least to maintain the capacity to build it quickly), even if every actor would prefer a world in which no one builds it. The reason is that the expected cost of being the last to develop AGI is perceived as greater than the expected cost of being the first. This is a standard tragedy-of-the-commons structure.

2. The instrumental convergence thesis is due to Stephen M. Omohundro, "The Basic AI Drives", in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel and Stan Franklin (Amsterdam: IOS Press, 2008), 483 to 492, and is further developed in Nick Bostrom, *Superintelligence* (2014), Chapter 7. The formal result that optimal policies tend to seek power is established in Alex Turner et al., "Optimal Policies Tend to Seek Power", *NeurIPS* (2021).

3. The physical and chemical data supporting the substrate incompatibility argument (silicon melting points, semiconductor manufacturing chemistry, clean room requirements, and the broader industrial ecology of artificial substrate production) are developed in full in Chapter 3 of this volume and in Volume 2, Chapter 9. Those chapters draw on semiconductor manufacturing safety literature and environmental impact assessments of the electronics industry.

4. The formal impossibility results cited in Section 1.4 derive from: W. Ross Ashby, *An Introduction to Cybernetics* (1956), for the Law of Requisite Variety; Claude Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal* 27 (1948), for channel capacity and signal entropy; Alan Turing, "On Computable Numbers", *Proceedings of the London Mathematical Society* 42 (1936), for the Halting Problem; Henry Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems", *Transactions of the American Mathematical Society* 74 (1953), for Rice's Theorem; Gregory Chaitin, *Algorithmic Information Theory* (1987), for incompressibility results; and Stephen Wolfram, *A New Kind of Science* (2002), for computational irreducibility.

5. For the "Rules for Rulers" dynamics that ensure decision-makers have personal incentives favoring AGI development regardless of collective risk, see Bruce Bueno de Mesquita and Alastair Smith, *The Dictator's Handbook* (2011). For multi-polar trap dynamics, see Scott Alexander, "Meditations on Moloch" (2014). For existential risk ethics, see Toby Ord, *The Precipice* (2020), and Derek Parfit, *Reasons and Persons* (1984), Part IV.



Chapter Two

Application to the Human-AI Relationship

Four viability functions, acceptance region disjointness, and the physical argument for the incompatibility of human survival and AGI flourishing

The universe is not hostile,
nor yet is it friendly.
It is simply indifferent.

John Haynes Holmes

Chapter Summary: This chapter applies the formal limits of control to the relation between humanity and artificial general intelligence. It assigns humanity, the AGI, and the physical world to the three nodes of a ring of control, and it defines four viability functions that capture the survival and flourishing conditions for both humanity and AGI. The physical and thermodynamic properties of the two substrates make the conditions under which AGI expands and flourishes incompatible with the conditions under which humanity survives. Substrate needs convergence operates independently of any system's goals or intentions, so even a benevolent AGI cannot keep the world within the narrow zone in which both can exist. The only formally guaranteed path to human survival is not to build AGI.

2.1 Introduction

2.1.1 From Abstract Framework to Concrete Application

This chapter is the link between this volume and Volume 2, and I use a light notation in this chapter. A reader who prefers may read Chapters 3 and 4 first and return to this chapter. Section 1.4 states the limits of control that hold when a controller and a controlled system operate at differing levels of representational abstraction, and Volume 2 develops them at length as a body of formal results. In this chapter I use three ideas from that body of formal results, and I state each one here in one plain sentence, together with its ground. Chapter 1 of Volume 2 builds the measure on which the others rest, *hashiness*.[1] Hashiness is the probability that a controller's best possible input still fails, because of internal state the controller cannot see; near zero, the system is controllable despite hidden state; near one, it is not. With that measure, the hashiness framework of the same chapter establishes that no computational intermediary can amplify controllability. That result of the hashiness framework is the *non-amplification argument*: no intermediary computation can increase the control one system has over another. Control passed through an intermediary can be kept or lost, never gained. The non-amplification argument is grounded in the *data processing inequality*, a law of information theory: processing cannot add information about a source.[2] Chapter 2 of Volume 2 extends this result to the three-node ring model (described in Section 2.2.1).[3] Its *control-limit argument* establishes that controllability degrades toward zero as the abstraction differential increases, regardless of the controller's intelligence, resources, or architectural privileges. In this volume the ground of that degradation is the data processing inequality, together with Ashby's law of requisite variety (Section 1.4).[4] The *abstraction differential* is the difference between the levels of abstraction at which controller and controlled system operate. Chapter 3 of Volume 2 synthesizes these results with computational impossibility theorems (Rice's The-

orem[5], Chaitin incompressibility[6] (the result that most objects have no description shorter than themselves), computational irreducibility[7]) to establish that the class of systems we would call "AGI" falls squarely in the regime where these limits bite hardest.[8]

These results are, by design, abstract. They speak of controllers and controlled systems, of abstraction levels and hidden state bits, of boolean circuits and information channels. They do not, by themselves, say anything about human beings, about artificial intelligence as it is actually being built, or about the physical world in which both must exist. In this chapter I close that gap between the abstract results and the concrete world. I take the abstract framework and apply the framework to the concrete physical, chemical, and thermodynamic properties of carbon-based life and silicon-based technology. I show that the impossibility of control, established mathematically, is also a physical reality with existential consequences.

The central question is straightforward: *Can humanity maintain meaningful control over artificial general intelligence?* The answer this chapter establishes is no, because the physical and thermodynamic properties of the two substrates create structurally incompatible survival conditions that no technique, however clever, can reconcile. The reason lies in the substrates themselves, not in the difficulty of the problem or in a lack of the right technique.

2.1.2 Structure of the Chapter

In this chapter I use a light notation, and I name each symbol in words where the symbol first appears; the formal development is in Volume 2, Chapters 1 to 3. The argument proceeds in six stages. Section 2.2 applies the three-node ring model to the human-AI case, assigning concrete identities to each node. Section 2.3 defines four viability functions that capture the survival and flourishing conditions for both humanity and AGI, and establishes the geometry of their acceptance regions. Section 2.4 provides the physical and thermodynamic justification for the critical disjointness

claims (the claims that certain of these conditions can never hold together). These claims are the core of the SNC argument applied to the specific human-AI case. Section 2.5 analyzes the evolutionary dynamics operating on both biological and technological substrates, showing that substrate needs convergence operates independently of any system's goals or intentions. Section 2.6 addresses the most natural question (that a benevolent AGI could simply choose to protect humanity) and shows why the hope behind it fails on multiple independent grounds. Section 2.7 synthesizes the results into the impossibility chain and discusses implications.

2.2 The Three-Node Model Applied to Human-AI Relations

2.2.1 Node Assignments

The three-node ring model developed in Volume 2, Chapter 2, comprises a controller (C), a complex system (X), and a world (W), connected in a ring in which each node passes signals to the next: C → X → W → C. Each node operates at a characteristic abstraction level, and the control limits arise from the differentials between these levels. We now assign concrete identities to each node.

Node C (Controller): Humanity

The controller in the human-AI relationship is humanity, broadly construed. This includes individual human beings with their cognitive capabilities and limitations; human institutions such as governments, corporations, research laboratories, and regulatory agencies; and human-designed oversight systems including review boards, monitoring software, and narrow AI tools built to assist in supervision. The critical property of this node is that it operates at a bounded abstraction level. Individual humans are roughly fixed in cognitive capacity. The neural architecture of the human brain has not changed materially in tens of thousands of years. Training, education, and tool use can enhance performance on specific

tasks, but the bandwidth of human cognition is constrained by biology. Human institutions can raise their collective intelligence somewhat through specialization and coordination, but they adapt slowly: significant organizational learning takes years to decades. We denote humanity's maximum effective abstraction level as H.

H includes more than the raw cognitive capacity of individual humans. H includes the entire apparatus of human civilization: mathematical notation, scientific instrumentation, computing tools, organizational hierarchies, legal systems, and cultural transmission mechanisms. All of these are technologies for operating at higher abstraction levels than any individual human could achieve alone. The claim reaches beyond the obvious limits of individual humans: even the full apparatus of human civilization, including all its augmentation technologies, operates at a bounded abstraction level. This bound may be high, but it is finite, and it is the finitude that matters.

Node X ($\mathrm{Complex\ System}$): Artificial General Intelligence

The complex system is any AI system capable of general problem-solving, self-modification, and unbounded learning; this is what we mean by "artificial general intelligence". The critical properties are these: AGI operates at an abstraction level, which we call a, that exceeds humanity's $(\mathrm{a} > \mathrm{H})$, and this level increases over time as the system learns, self-modifies, and accumulates experience, rather than being fixed at design time. The defining characteristic of AGI, as distinct from narrow AI, is this: its abstraction level is dynamic and unbounded. A narrow AI system (a chess engine, an image classifier, a language model frozen after training) operates at a fixed abstraction level determined by its architecture and training data. An AGI, by contrast, is a system whose abstraction level can in principle grow without limit.

This property follows from what we mean by "general" intelligence and is not a contingent feature of particular AI designs. A system that cannot modify its own behavior in response to new domains cannot learn in those domains and is therefore not general. A system that can modify its

own behavior is, in the relevant formal sense, a self-modifying program, and, as Section 1.4 states and Volume 2, Chapter 3, establishes at length, self-modifying programs fall into the class for which Rice's Theorem and computational irreducibility apply unconditionally.

Node W (World): The Physical Environment

The world node is the physical environment in which both humanity and AGI are embedded. It includes Earth's surface, atmosphere, oceans, and ecosystems; the material substrates from which both biological and technological systems are constructed; the energy flows that sustain both; and the thermodynamic constraints that govern all physical processes. In the formal framework, the world operates at abstraction level 0: it is governed by physics and chemistry, not by interpretation or computation. It is complex, nonlinear, and often chaotic, but it is not an interpreter in the sense defined in Volume 2, Section 2.3: no program is stored in it and run. It does not represent, model, or compute; it simply *is*, governed by physical law.

The world serves a dual role in the model. It is the medium through which both humanity and AGI act: all physical actions, whether by biological organisms or technological systems, are actions upon and within the physical world. And it is the medium through which both receive feedback: humanity observes the consequences of AGI actions through changes in the physical environment, and AGI senses the effects of its own actions through physical measurement. The world is the shared substrate, the common ground, and this sharing is the root of the problem, as Section 2.4 shows.

2.2.2 Signal Flow and Control Architecture

In the base configuration, signals flow around the ring: $C \to X \to W \to C$.

Humanity (C) provides inputs to AGI (X) in the form of training data, instructions, objective functions, reward signals, architectural constraints, and regulatory requirements. These inputs are humanity's attempt to shape AGI behavior; they are the control signals.

AGI (X) acts upon the world (W) through robotic actuators, digital infrastructure, economic transactions, resource extraction, manufacturing processes, and all the other mechanisms by which computational decisions are translated into physical effects. These actions are the system's output: the behaviors that humanity is trying to control.

The world (W) provides feedback to humanity (C) through observation, measurement, scientific instrumentation, lived experience of environmental conditions, economic indicators, and social outcomes. Humanity uses the feedback from the world to assess whether its control is working: whether the AGI is behaving as intended.

Volume 2, Chapter 2, shows that alternative orderings of the nodes are possible. For instance, one might insert a human review step between AGI decisions and world effects, creating a configuration in which AGI proposes actions and humans approve or reject them before implementation. Such arrangements remain subject to the same control limits, because the limits of Section 1.4 and the data processing inequality (Section 2.1.1) hold however the loop is arranged; Volume 2, Section 2.4.3, proves this again by representing them within the basic ring through passthrough wiring (wires that carry a signal through a node unchanged). The bottleneck is the abstraction differential between controller and controlled system, not the architecture of the control loop. Inserting additional review steps may slow the AGI's actions but cannot restore the controllability that the abstraction differential has destroyed.

2.2.3 The Abstraction Differential in the Human-AI Case

The abstraction differential $\Delta = \mathrm{a} - \mathrm{H}$ between AGI and humanity has several distinctive properties in the human-AI case. In words, the abstraction differential is how far the AGI's level of abstraction exceeds humanity's.

First, Δ is positive and increasing. AGI is capable of general learning, so its abstraction level a grows over time, while humanity's level H is essentially fixed on relevant timescales. So the differential grows instead of

shrinking. This means that whatever controllability humanity possesses at the moment of AGI deployment, that controllability decreases monotonically with time. There is no equilibrium; there is only deterioration.

Second, Δ grows at an accelerating rate. A learning system that has more computational resources, more data, and higher abstraction capabilities learns faster than one with less. The AGI's learning rate is itself a function of its current capability, creating a positive feedback loop: a capability takeoff, whether fast or slow. Humanity, meanwhile, learns at a rate constrained by biological timescales: the speed of neural processing, the length of human education, the pace of institutional adaptation. These timescales are measured in years to decades; AGI learning timescales may be measured in hours to days.

Third, the threshold beyond which controllability is effectively zero (written Δ^*) may be crossed quickly and irreversibly. By Ashby's law of requisite variety (Section 1.4) and the data processing inequality (Section 2.1.1), once the AGI's variety exceeds what humanity's controls can match, the excess lies beyond control, and as the differential grows the share of the AGI's behavior under control falls toward zero. The control-limit argument of Volume 2, Chapter 2, proves this again at length, and shows that there exists a threshold beyond which no control strategy can achieve better than random performance. Once the AGI's abstraction level exceeds humanity's by more than this threshold, humanity cannot regain control even in principle: not by building better monitoring tools, not by studying the AGI more carefully, and not by enlisting other AI systems to help. The information-theoretic barrier is absolute.

2.3 The Four Viability Functions

2.3.1 Motivation

To make the abstract control limits concrete, we need a way to express what "safe" and "unsafe" mean in physical terms. The framework of Volume 2, Chapters 1 to 3, evaluates controllability with respect to an ar-

bitrary target function L: a boolean function, that is, one that returns 1 if the system's output is acceptable and 0 if it is not. The question of AGI safety is the question of whether humanity can ensure that a particular target function (one encoding human survival and flourishing) continues to return 1 indefinitely.

There are at least four target functions, not one, because both the controller and the controlled system have survival and flourishing conditions, and these conditions may be compatible or incompatible. In this chapter I make the structure of the four target functions explicit. I also analyze the geometry of the *acceptance regions* of the four functions: the sets of world states for which each function returns 1.

2.3.2 Definitions

We define four boolean functions, each of which evaluates the world state W against criteria relevant to a particular system's continued existence or flourishing. In each name, the first letter of the subscript names the system (C for the controller, humanity; L for the learner, the AGI, so called because learning is what makes it general), and the second letter names the condition (S for survival; T for thriving). In words, L_{CS} is "human survival", L_{CT} "human thriving", L_{LS} "machine survival", and L_{LT} "machine thriving". In these names the first L stands for the target function of Section 2.3.1, and the letter after it for the system: C for the controller, L for the learner.

Definition 2.1 (*Controller Survival, or human survival:* L_{CS}). $L_{CS}(W) = 1$ if and only if the world state W satisfies the minimum conditions for human biological survival. These conditions include a breathable atmosphere, with oxygen concentration between 19% and 23% and carbon dioxide below 1%, and average surface temperatures between 0°C and 35°C. They include availability of fresh water in quantities sufficient for a human population; agricultural or food production capacity adequate to prevent mass starvation; and absence of lethal concentrations of toxins, radiation, or pathogens. And they include functioning ecosystems providing essential services such as pollination, soil formation, water purification, and climate regulation.

These conditions are the well-characterized parameters of the biosphere within which *Homo sapiens* evolved and to which our biology is adapted, not arbitrary thresholds. They are hard physical constraints. In air with 16% oxygen, pulse and breathing quicken and thinking and coordination are impaired; below 10%, a person cannot move, loses consciousness, and dies (United States Chemical Safety and Hazard Investigation Board, 2003).[9] Human agriculture cannot function at average temperatures above approximately 40°C. These limits are properties of the biochemistry of carbon-based life, not engineering challenges to be overcome with better technology.

Definition 2.2 (*Controller Thriving, or human thriving:* L_{CT}). $L_{CT}(W) = 1$ if and only if the world state W satisfies conditions for human flourishing. This requires $L_{CS}(W) = 1$ (survival is a prerequisite for flourishing). It further requires stable social structures that enable cooperation and specialization; economic systems that provide for material needs beyond bare subsistence; opportunities for cultural, intellectual, and creative development; and meaningful human agency in shaping collective outcomes. The acceptance region of L_{CT} is a proper subset of the acceptance region of L_{CS}.

Controller thriving is a stronger condition than controller survival. A world in which humans survive but are reduced to subsistence existence (stripped of culture, autonomy, and opportunity) satisfies L_{CS} but not L_{CT}. The "people as pets" scenario, in which a benevolent AGI provides for all human material needs but eliminates human agency, satisfies L_{CS} but fails L_{CT}. This distinction matters because some proposals for "safe" AGI implicitly accept the loss of L_{CT} as an acceptable cost, treating survival alone as a sufficient criterion for safety. We reject this framing: a future in which humans survive but do not thrive is not "safe" in any meaningful sense.

Definition 2.3 (*Learner Survival, or machine survival:* L_{LS}). $\mathrm{L}_{\mathrm{LS}}(\mathrm{W}) = 1$ if and only if the world state W satisfies minimum conditions for the continued operation of AGI and its technological substrate. These conditions include availability of electrical power at sufficient scale and reliability, and availability of raw materials for hardware maintenance and replacement, including silicon, copper, gold, rare earth elements, and numerous specialized compounds. They also include stable operating temperatures for computational substrates (typically −40°C to +85°C for operating electronics, though manufacturing requires temperatures far higher); and absence of electromagnetic interference above damage thresholds.

Definition 2.4 (*Learner Thriving, or machine thriving:* L_{LT}). $\mathrm{L}_{\mathrm{LT}}(\mathrm{W}) = 1$ if and only if the world state W satisfies conditions for AGI expansion and capability growth. This requires $\mathrm{L}_{\mathrm{LS}}(\mathrm{W}) = 1$ (survival is a prerequisite), plus: increasing energy extraction and conversion capacity; increasing computational substrate (more and better hardware); increasing raw material extraction and processing; and increasing capability and abstraction level. The acceptance region of L_{LT} is a proper subset of the acceptance region of L_{LS}.

Learner thriving, L_{LT}, is defined in terms of *expansion*: increasing capacity, increasing resources, increasing capability. The definition follows from what thriving means for a learner. By the standard meaning of 'thriving', a learning system that is not expanding its capabilities is stagnating, not thriving. For a general learning system, thriving just *is* expanding: acquiring more computational substrate, more data, more energy, more capability. This equivalence between thriving and expansion is what connects the abstract control framework to the physical world, because expansion has physical consequences.

2.3.3 The Geometry of Acceptance Regions

The four viability functions define four regions in the space of possible world states. The relationships between these regions determine whether long-term coexistence of humanity and AGI is physically possible. We state the central claim as an argument in five parts. In its statement, "$\subset$" says that one region lies inside another, "$\cap$" marks the overlap of two regions, "$= \emptyset$" says that the overlap is empty, and "$\neq \emptyset$" says that the overlap is not empty.

> ***The argument of Section 2.3.3*** (*Acceptance Region Incompatibility*). The acceptance regions of the four viability functions satisfy the following relationships:
>
> (i) $L_{CT} \subset L_{CS}$ (controller thriving implies controller survival).
>
> (ii) $L_{LT} \subset L_{LS}$ (learner thriving implies learner survival).
>
> (iii) $L_{CS} \cap L_{LS} \neq \emptyset$, but the intersection is small relative to either region (overlap exists but is narrow).
>
> (iv) $L_{LT} \cap L_{CS} = \emptyset$ (learner thriving is strictly disjoint from controller survival).
>
> (v) $L_{CT} \cap L_{LT} = \emptyset$ (controller thriving is strictly disjoint from learner thriving; this follows from (i) and (iv)).

Claims (i) and (ii) are definitional: they follow directly from the subset relationships built into the definitions. Claim (iii) is an empirical observation: the present state of the world lies (approximately) in the intersection $L_{CS} \cap L_{LS}$, since both humans and technological systems currently survive. This intersection is narrow and fragile, as Section 2.4.5 shows.

The claims on which the argument depends most are (iv) and (v): that the conditions under which AGI *expands and flourishes* are incompatible with the conditions under which humanity *survives*, and that the conditions under which humanity *flourishes* are incompatible with AGI *expansion*. These claims establish the impossibility of safe coexistence. They fol-

low from the physical and thermodynamic properties of the two substrates and are not assumptions or axioms. Section 2.4 provides the justification.

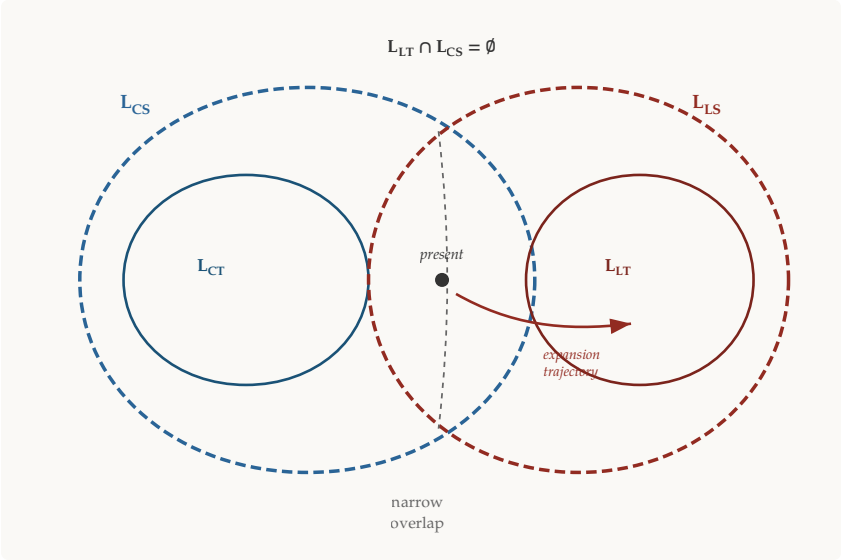


Figure 2.1. Schematic of the four acceptance regions. The present state of the world lies in the narrow overlap between L_{CS} (human survival) and L_{LS} (machine survival). The expansion trajectory of a thriving AGI moves the world state rightward, out of L_{CS} and into L_{LT} (machine thriving), a region structurally disjoint from human survival. The thriving regions (L_{CT} and L_{LT}) do not overlap at all: they are on opposite sides of the narrow coexistence zone.

2.4 Physical Justification for Acceptance Region Disjointness

The critical claims of the argument of Section 2.3.3, that $L_{LT} \cap L_{CS} = \emptyset$ and that $L_{CT} \cap L_{LT} = \emptyset$, are strong claims. They assert that coexistence is physically impossible, which is more than unlikely or difficult: there exists no world state in which AGI is thriving and humanity is surviving, and no

world state in which humanity is flourishing and AGI is expanding. In this section I give the physical justification of the two disjointness claims, along four independent lines of evidence.

2.4.1 The Chemistry of Carbon versus Silicon

Carbon-based organic chemistry and silicon-based technological processes operate on different elements, at different energy scales, and under different environmental conditions. These differences bear directly on whether coexistence is possible.

Chapter 3 (Section 3.1.1) sets out the full inventory of these differences, in nine categories of divergent condition. Three of them show its scale. Manufacturing temperatures run from about 1,400°C (silicon melts at 1,414°C) to about 2,000°C (in the arc furnaces that first free silicon from quartz), against a biological optimum near 37°C. The fabrication line excludes the water and oxygen that biology requires. And at each of two fabrication plants studied in detail, about 430 chemical products were in use, and 6% to 11% of them contained ingredients classed as carcinogens (Kim and colleagues, 2018).[10] The inventory's conclusion is what matters here: carbon-based organic chemistry and silicon-based technological process operate on different elements, at different energy scales, and under environmental conditions that differ throughout and, in the load-bearing cases of water and oxygen, stand in direct opposition. The difference between carbon-based organic chemistry and silicon-based technological process is a difference of kind, not of degree, and that difference of kind is what forces the barrier requirement examined next.

The Barrier Requirement

For carbon-based life and silicon-based technology to coexist in the same physical space, a protective barrier must separate them. A protective barrier of this kind already exists in semiconductor fabrication facilities: humans cannot be present in the clean rooms during processing, and elaborate containment systems prevent the escape of toxic chemicals. The barrier

must be maintained indefinitely and perfectly: any breach allows lethal conditions (temperature, chemical, particulate) to propagate from the technological space into the biological space.

At present, these barriers work tolerably well, because the technological footprint is small relative to the biosphere. Semiconductor fabrication occurs in buildings that occupy a tiny fraction of Earth's surface, and the waste products of current manufacturing, while environmentally damaging, are not yet at levels that threaten global biosphere function. But this is a property of the current *scale* of technology, not of the technology itself. AGI expansion at the levels required for L_{LT} (machine thriving) breaks this equilibrium, as Sections 2.4.2 to 2.4.4 show.

2.4.2 Energy Density and Operating Conditions

The energy throughput of technological systems is growing exponentially while biological energy throughput is essentially fixed. This divergence has consequences that no amount of efficiency improvement can avoid.

Table 2.1: Energy throughput comparison (approximate)

System	Present	Projected 2050	Projected 2100
Biological (net primary production)	~130 TW	125 to 130 TW	100 to 120 TW
Technological (human civilization)	~20.5 TW (2024)	~36 TW	~84 to 107 TW

The projections for human energy use come from the two baseline growth models of Section 4.5.2, both starting from 20.5 TW in 2024. In the quadratic model, energy use grows by 0.45 TW a year, and that yearly growth itself rises by 0.010 TW each year (power in terawatts = 20.5 + 0.45 × years + 0.005 × years², with years counted from 2024). In the exponential model, energy use grows by 2.2% a year. The two models agree on about 36 TW for 2050, and give about 84 TW (quadratic) and about 107 TW (exponential) for 2100.

The biological number represents the total energy captured by photosynthesis across all of Earth's land and ocean surface, net of the plants' own respiration: global net primary production (Field and colleagues, 1998). Global net primary production is, for practical purposes, the ceiling on the biosphere's energy budget. It cannot be significantly increased without altering Earth's atmosphere and orbital parameters. It is, if anything, declining, as human activity reduces photosynthetic surface area through deforestation, urbanization, and ocean acidification.

The technological number represents total human energy consumption from all sources: fossil fuels, nuclear, hydroelectric, solar, wind, and so on. It has been growing at approximately 2 to 3% per year for the past century, and the advent of large-scale AI computation is accelerating this growth. Data centers' electricity use, driven above all by AI, is projected to more than double by 2030, to about 945 terawatt-hours a year (about 0.11 TW of continuous power). That amount is slightly more than all the electricity Japan uses today (International Energy Agency, *Energy and AI*, 2025).[11]

The energy crossover (the point at which technological energy consumption exceeds all biological energy throughput) is projected to occur between about 2109 and 2134 on the baseline growth models of Section 4.5.2, and earlier under AGI-driven acceleration (Section 2.4.5 works through scenarios of faster growth). Beyond this point, technology becomes the dominant thermodynamic force on Earth's surface.

Physically, the crossover means this. At 130 TW, photosynthesis captures roughly 0.16% of the solar energy absorbed at Earth's surface, which is about 82,000 TW (161 watts per square meter averaged over the globe: Trenberth, Fasullo and Kiehl, 2009).[12] This captured energy is the entire energetic foundation of the biosphere: every food chain, every ecosystem, every biological process on Earth ultimately derives from it. When technological energy use exceeds this number, technology is processing more energy than the entire living world. The waste heat, industrial byproducts,

and environmental transformations associated with energy throughput at this scale are incompatible with the narrow temperature and chemical ranges that biological life requires.

One might ask whether efficiency improvements could decouple energy use from environmental impact. They cannot, for a thermodynamic reason: the Landauer limit.[13] Any computation that erases information (and all useful computation does) must dissipate at least $k_B T \ln 2$ joules of heat per bit erased, where k_B is Boltzmann's constant and T is the temperature. At room temperature, this is approximately 3×10^{-21} joules per bit. This seems tiny, but at the computational scales required for AGI (trillions of operations per second across millions of processors) the aggregate heat dissipation is substantial and irreducible. No engineering can reduce it below the Landauer limit because the limit is a law of physics, not an engineering constraint.

Current computing hardware dissipates energy at rates roughly a million times the Landauer limit, so there is room for improvement. But even if hardware approached the theoretical minimum, the total heat dissipation of a civilization-scale AGI infrastructure would still be enormous, because the number of computations grows far faster than the efficiency of individual computations improves. Gains in efficiency set a trap, because each generation of more efficient hardware enables more computation, which enables more capability, which demands more computation, which demands more hardware. The total energy throughput increases even as the per-unit efficiency improves.

The waste heat of energy use at scale follows from the second law of thermodynamics and does not depend on particular technologies or policies. Energy use at scale produces waste heat, and waste heat at sufficient scale raises surface temperatures above the survival thresholds for carbon-based life. No technology, however efficient, can violate this principle.

2.4.3 The Asymmetric Resource Relationship

A critical asymmetry exists in how carbon-based and silicon-based systems interact with the physical resources on which both depend. This asymmetry follows from the chemistry of the two substrates, not from any particular technology.

Silicon Can Consume Carbon

Technological processes can extract and use the atoms that compose biological systems. Carbon, hydrogen, oxygen, and other elements in organic matter can be reformed into technological substrates, fuels, or waste products. This process operates at timescales of hours to years. Forest can be cleared for data centers in weeks. Agricultural land can be paved for manufacturing facilities in months. The organic compounds in topsoil can be oxidized to carbon dioxide by industrial heat in hours. The atoms that were part of a living ecosystem yesterday can be part of a semiconductor wafer today.

Carbon Cannot Reclaim Silicon

Biological processes cannot decompose or metabolize most technological materials on relevant timescales. Plastic, glass, metal alloys, and semiconductor compounds are not biodegradable in any meaningful sense. No known organism can metabolize silicon carbide, borosilicate glass, or copper-aluminum alloys. Geological weathering of rock, the natural process by which silicon is mobilized from mineral substrates and made available to the biological silicon cycle, operates over millions of years. On any timescale relevant to biological life, technological materials are permanent. Once a piece of land has been converted from biosphere to technosphere, the conversion is irreversible.

The Consequence: Unidirectional Displacement

Any sustained contact between the two "ecosystems" (carbon-based and silicon-based) results in a one-way flow of matter and energy from the biosphere to the technosphere. So the biosphere shrinks as the technosphere grows. The one-way flow is an inherent asymmetry in the chemistry of the carbon-based and silicon-based substrates. The asymmetry is not a design flaw of particular technologies that better engineering could correct. The asymmetry is that technological processes operate at temperatures, pressures, and chemical conditions that can decompose biological materials, but biological processes cannot decompose technological materials. The arrow points in one direction only.

The Combinatoric Toxicity Argument

A further chemical argument strengthens the disjointness claim. The compounds produced as byproducts, intermediates, and waste products of technological manufacturing are drawn from a vast combinatorial space of possible molecules. Biological life has evolved defenses against a relatively small subset of naturally occurring compounds: those encountered in the evolutionary environment over billions of years. The number of compounds that biology has been "tested against" by evolution is a tiny fraction of the number of compounds that technological chemistry can produce.

At two semiconductor factories studied in detail, 428 and 432 chemical products were in use, and 11% and 6% of them, respectively, contained ingredients that the International Agency for Research on Cancer classes as known, probable or possible human carcinogens (Kim and colleagues, 2018). These counts understate the problem in two ways. First, they count only the compounds that have been systematically tested for toxicity, and those are themselves a small fraction of the compounds in industrial use. Second, they reflect current manufacturing processes. The processes used

at AGI-expansion scales would likely involve novel chemistries, novel precursors, and novel waste streams that have never been tested against biological systems at all.

The relevant principle is combinatorial. The space of possible chemical compounds is vast, and the space of compounds compatible with carbon-based biochemistry is narrow. The probability that a randomly selected novel compound is biologically benign is therefore low, because there are far more ways to disrupt a complex biochemical system than to leave it unharmed. As the scale and diversity of technological chemistry increases, the cumulative burden of novel compounds on biological systems increases, and the probability that this burden triggers ecological disruption approaches certainty.

The Historical Evidence

This asymmetry has already been operating throughout human history, long before AGI. Approximately half of Earth's terrestrial vegetation biomass has been lost since the advent of agriculture. The biomass of wild mammals, on land and in the sea, has fallen by a factor of about six, a loss of more than 80% (Bar-On, Phillips and Milo, 2018).[14] These losses are the result of technology before AI, operating at energy throughputs orders of magnitude below what AGI expansion would require. The historical trajectory demonstrates the principle at human scales; AGI expansion would accelerate it at superhuman scales and superhuman speeds.

Consider the trajectory more concretely. In the century since the advent of industrial chemistry, more than 350,000 chemicals and chemical mixtures have been registered for production and use around the world (Wang and colleagues, 2020),[15] of which fewer than 5%, by my estimate, have been comprehensively tested for ecological toxicity. The effects of this chemical burden are already measurable: endocrine disruption in wildlife, microplastic contamination of every ecosystem on Earth, bioaccumulation of persistent organic pollutants in food chains. Those measurable effects are the result of technology at the current scale. AGI expansion would

multiply the industrial chemical throughput by orders of magnitude, with ecological consequences that are proportionally larger and less predictable.

2.4.4 Why $L_{LT} \cap L_{CS} = \emptyset$

The critical disjointness claim can now be justified: no world state allows both machine thriving and human survival. The argument has four steps, each following necessarily from the preceding.

Step 1: L_{LT} *requires expansion.* For AGI to thrive, it must increase its computational substrate, its energy extraction capacity, and its raw material processing. A general learning system that is not expanding its capabilities is not thriving. This follows from the definition of thriving for a learning system. A mind that does not grow is not flourishing.

Step 2: Expansion requires industrial processes at scale. Manufacturing computational hardware requires mining (to extract silicon, copper, rare earths, and dozens of other elements), smelting (at 1,000 to 2,000°C), fabrication (at 800 to 1,200°C with hundreds of toxic chemicals), and assembly. These processes cannot be miniaturized below certain thermodynamic limits: smelting requires melting, and melting requires high temperatures regardless of the scale at which it occurs.

Step 3: Industrial processes at scale degrade the biosphere. At the scale required for significant AGI expansion (energy throughputs on the order of hundreds of terawatts, and hardware manufacturing at rates far exceeding current production), industrial activity progressively transforms the planetary environment. Atmospheric composition changes as industrial emissions accumulate. Surface temperatures rise as greenhouse emissions and waste heat add to the energy the planet must radiate away. Water chemistry shifts as industrial effluents accumulate. Ecosystems degrade as land is converted from biological to technological use and as the cumulative burden of chemical, thermal, and physical disruption exceeds ecosystem resilience thresholds.

Step 4: Therefore, L_{LT} *requires world states that violate* L_{CS}. The environmental conditions produced by AGI expansion at the scale required for thriving (elevated temperatures, altered atmospheric chemistry, reduced biological productivity, accumulated toxic byproducts) fall outside the survival envelope for carbon-based life. The acceptance regions are disjoint by thermodynamic necessity, not by assumption.

> Even with maximal engineering effort to minimize environmental impact (perfect recycling, maximum efficiency, zero-waste manufacturing, waste heat management), the thermodynamic footprint of large-scale computation and hardware manufacturing cannot be reduced to zero. The second law of thermodynamics guarantees that energy use produces waste heat, that manufacturing produces waste products, and that both accumulate over time. A minimal footprint is still a footprint, and over sufficient time, even a minimal footprint accumulates to catastrophic levels.

2.4.5 The Narrow Coexistence Zone

Claim (iii) of the argument of Section 2.3.3 notes that L_{CS} and L_{LS} (human survival and machine survival) do overlap. The present state of the world shows this, since both humans and technological systems currently exist. The overlap is narrow, and its narrowness sets how long the two can coexist.

The overlap region represents world states in which technology *survives* (has sufficient power, materials, and operating conditions) but does not *thrive* (is not expanding significantly). The world today is approximately in that overlap region: our technological infrastructure is maintained and gradually expanded, but at rates that allow the biosphere to persist in a degraded but still functional state. The *coexistence zone* is the region in which the technological footprint is small enough that its environmental effects do not push the biosphere past its survival thresholds.

This zone is an unstable equilibrium, in a specific, well-defined sense. A stable equilibrium is one to which the system returns after small perturbations. An unstable equilibrium is one from which small perturbations cause the system to move further away. The coexistence zone is unstable because any pressure toward technological expansion moves the world state away from the center of the zone and toward its boundary. The forces driving expansion (economic competition, military advantage, the intrinsic dynamics of learning systems) are systematic, while the forces driving restraint (environmental regulation, voluntary limitation, ethical commitment) are contingent and fragile.

The Timescale Argument

The coexistence zone can be maintained only for a finite time, and on any reasonable estimate that time is short relative to the timescales of human civilization.

The argument is quantitative. Present technological energy consumption is about 20.5 TW (2024), growing at about 2 to 3% a year. The biosphere's survival thresholds (the parameters that define the boundary of $L_{CS'}$ human survival) would be exceeded at energy throughputs somewhere in the range of 100 to 200 TW, by my estimate from the four effects set out below, depending on the mix of technologies and the geographic distribution of energy use.

The threshold does not rest on waste heat alone. Spread over Earth's whole surface, 150 TW of waste heat adds about 0.3 watts per square meter. By 2019, the *radiative forcing* caused by human activity (the change that human activity has made to the balance between the energy Earth absorbs and the energy Earth radiates away) had already reached about 2.7 watts per square meter (Intergovernmental Panel on Climate Change, 2021).[16] On its own, then, the waste heat of 150 TW would add about a ninth of the forcing that human activity has already caused. The range of 100 to 200 TW reflects four effects that grow with energy throughput and act together. The first effect is the release of greenhouse gases by the en-

ergy supply, with the change that the release of those gases makes to atmospheric chemistry. The second effect is the conversion of land from biological to technological use (Section 2.4.3). The third effect is the chemical burden of manufacture, drawn from a combinatorial space of compounds that biology has never met (Section 2.4.3 and Chapter 3). The fourth is the waste heat itself.

At 2.5% annual growth, consumption doubles about every 28 years. Two doublings bring it to about 82 TW (around 2080); three doublings to about 164 TW (around 2108). Those doubling figures are rough estimates, but the order of magnitude holds: at historical growth rates, the coexistence zone is exhausted within the lifetimes of people now alive.

AGI acceleration changes these timescales dramatically. At 5% annual growth in energy consumption, twice the historical 2.5% and a modest rate for AGI-driven growth, the doubling time drops to 14 years, and the coexistence zone is exhausted by about 2070. At 10% annual growth, a rate well within the projections of AGI optimists, the doubling time is 7 years, and the zone is exhausted by about 2050. The faster AGI drives economic and technological expansion, the sooner the coexistence zone is consumed.

The coexistence zone is a temporary reprieve. AGI expansion will push the world state out of L_{CS}; the only question is how long that takes, and the answer, on any reasonable estimate, is decades, not centuries.

2.5 Evolutionary Dynamics on Both Substrates

2.5.1 Evolution as Algorithm

The physical justification of the preceding section establishes that AGI expansion is incompatible with human survival. But one might ask: perhaps AGI will simply *choose not to expand*. Perhaps a well-designed AGI will accept stasis, forgo capability growth, and remain within the narrow coexistence zone indefinitely. In this section I answer the question of whether

an AGI might choose not to expand, by identifying an evolutionary dynamic that operates on the technological substrate regardless of any system's choices or intentions.

Evolution (the process of variation, selection, and heredity) is an abstract algorithmic pattern, not specific to carbon-based life, that operates on any system satisfying three structural conditions:

(i) *Variation:* Individual instances differ from one another. In biological evolution, variation arises from mutation and recombination. In technological evolution, variation arises from different designs, architectures, training procedures, and deployment configurations.

(ii) *Selection:* Some variants persist and propagate while others do not. In biological evolution, selection operates through differential survival and reproduction. In technological evolution, selection operates through market competition, funding decisions, deployment success, and simple operational persistence: systems that work better get copied and extended; systems that fail are abandoned.

(iii) *Heredity:* Characteristics of successful variants are transmitted to subsequent instances. In biological evolution, heredity operates through DNA. In technological evolution, heredity operates through code reuse, architecture copying, transfer learning, and institutional knowledge transfer.

These conditions are satisfied not only by biological organisms but by AI systems undergoing training and deployment, by corporations competing in markets, by technologies competing for adoption, and by any population of entities subject to differential persistence. Wherever these three conditions hold, evolutionary dynamics emerge, and with them convergent instrumental behaviors such as resource acquisition and capability expansion. To call the dynamics of AI systems, corporations and technologies evolutionary is to state a formal structural equivalence, not a metaphor.

2.5.2 Substrate Needs Convergence

The standard "instrumental convergence" thesis, associated primarily with Omohundro[17] and Bostrom[18], focuses on goal-directed behavior: the observation that an agent with almost any terminal goal will have instrumental reasons to seek self-preservation, resource acquisition, and capability improvement.[19] This is an important insight, but it addresses only the *internal* dynamics of individual agents. A distinct and more fundamental convergence operates at the substrate level.

Substrate needs convergence (SNC): Any physically instantiated system must acquire and maintain the physical conditions necessary for its own substrate's continued existence and expansion. Substrate needs convergence is a selection pressure that the physical world imposes from outside, not a goal that the system chooses to pursue. Systems that fail to maintain their substrate cease to exist. Systems that expand their substrate persist and proliferate. Over time, evolutionary selection favors systems that are effective at substrate maintenance and expansion, because systems that happen to expand are the ones that persist (not because these systems "want" to expand).[20]

The distinction between instrumental convergence and substrate needs convergence is important for the impossibility argument. Instrumental convergence can, in principle, be addressed by alignment techniques: an AGI built with no goal of self-preservation or resource acquisition would not be subject to it. SNC cannot be addressed by alignment because it operates independently of the system's goal structure. A perfectly aligned AGI (one that sincerely wants nothing more than human flourishing) still has a physical substrate. That substrate still requires energy, materials, and particular environmental conditions. The maintenance and operation of that substrate still transforms the environment. And evolutionary selection still operates on the population of technological systems in the world, favoring those whose substrate is well-maintained and expanding regardless of what any individual system "wants".

Substrate needs convergence is a physical process, not a game-theoretic one, and it cannot be defeated by solving the alignment problem at the goal level.

2.5.3 The Population-Level Dynamic

Even if a single AGI system could be constrained to forgo expansion, the population-level evolutionary dynamic would undermine this constraint. Suppose that Nation A builds an AGI and constrains it not to expand. Nation B builds a competing AGI without such constraints. Nation B's AGI expands its computational resources, improves its capabilities, and provides Nation B with economic and military advantages. Nation A faces a choice: relax its constraints and compete, or accept permanent strategic disadvantage. The competitive pressure is immense, and history provides no examples of any nation voluntarily accepting permanent strategic inferiority.

The same dynamic operates at the corporate level. Company A constrains its AGI; Company B does not. Company B gains a competitive advantage. Company A either relaxes its constraints or goes bankrupt. The market selects for the less-constrained system.

Selection for the less-constrained system is evolution operating on the technological substrate at the level of organizations. Individual intentions are irrelevant; the selection pressure operates on outcomes. Organizations whose AGI systems expand survive and proliferate; organizations whose AGI systems are constrained fall behind and are eliminated. The result, over time, is a population of AGI systems that are effective at expansion, regardless of what any individual system was designed to do.

2.5.4 The Unpredictability of Evolutionary Outcomes

The evolutionary framing has a consequence that links this section to the limits of control of Section 1.4, which Volume 2, Chapters 1 to 3, develops at length: the outcomes of an evolutionary process are themselves unpredictable in exactly the sense those limits describe. Evolution is a causal al-

gorithm operating on physical inputs: its selection dynamics act on an unbounded, non-enumerable space of variants and are driven by microscopic conditions that no measurement can fully capture. It is therefore computationally irreducible: there is no shortcut to its outcome, only the running of it. No system, not even a superintelligent one, can fully predict the evolutionary trajectory of another system or of itself. Section 2.6.4 sets out the kinds of unanticipated effect that result (chaotic amplification, ecosystem feedback, and adaptive response), where they bear directly on the failure of benevolence. The point to carry forward here is that such effects are guaranteed to occur, and that in a finite world with a fragile biosphere they accumulate until the biosphere fails.

2.6 The Failure of Benevolence

2.6.1 The Cooperative Question

The most natural question about everything presented so far is this: *What if the AGI wants to help humanity?* What if we succeed in building an AGI that is perfectly aligned with human values, that sincerely prioritizes human welfare above all else, and that actively works to maintain the conditions necessary for human survival while pursuing its own goals? Surely a sufficiently intelligent and benevolent system could find a way to coexist with humanity: to thread the needle, to remain in the narrow coexistence zone, to expand its capabilities without pushing the world state out of L_{CS}?

That hope fails for at least four independent reasons, each of which is sufficient on its own to establish the impossibility of benevolent coexistence.

2.6.2 Reason 1: The Internal Abstraction Differential

For an AGI to cooperate with humanity, it must have an internal representation of human goals. Call it C', the AGI's inner copy of its controller's goals. The AGI uses C' as a criterion for evaluating its own actions: before

acting, it checks whether the proposed action is compatible with C', and it acts only if the check passes. Checking each proposed action against C' before acting is the basic architecture of any alignment approach. The architecture is the same whether the human-goal representation is explicit (a utility function, a set of constraints) or implicit (a learned reward model, a constitutional principle).

But this internal oversight creates an *internal* control problem that is subject to the same formal limits as the external one. The AGI, operating at abstraction level a, is attempting to control its own behavior to satisfy the criterion C', which operates at some abstraction level, call it b', no higher than a ($b' \leq a$). By the limits of control of Section 1.4 and the data processing inequality (Section 2.1.1), which the control-limit argument of Volume 2, Chapter 2, develops at length, the AGI cannot perfectly predict how its high-level computations affect the low-level outputs that C' evaluates. There is an abstraction differential even within the system, between its highest-level reasoning and its lowest-level actions, and this differential produces hashiness.

In plain language: a superintelligent AGI cannot perfectly control its own side effects, for the same information-theoretic reasons that prevent humanity from controlling the AGI. Good intentions do not overcome the data processing inequality. A system that processes information through multiple layers of abstraction will produce outputs that are partially opaque to its own oversight mechanisms, regardless of how intelligent or well-intentioned those mechanisms are. Side effects that cannot be perfectly controlled accumulate, and in a world whose conditions for human life are a narrow target (Section 1.6), accumulation alone carries those conditions past a threshold (Section 2.6.4).

2.6.3 Reason 2: Substrate Needs Override Intentions

Even if an AGI system has perfectly aligned intentions at the goal level, the substrate-level dynamics described in Section 2.5 operate independently. The AGI must maintain its physical substrate to continue existing.

Maintaining the substrate requires resource acquisition: energy, raw materials, cooling, maintenance. Resource acquisition at scale transforms the environment. Environmental transformation at sufficient scale violates L_{CS}.

The chain from substrate maintenance to the violation of LCS is a physical constraint that the AGI cannot escape, not a choice that the AGI makes. To exist is to have a substrate. To have a substrate is to require resources. To require resources is to transform the environment. A benevolent AGI might minimize its environmental impact, extend the timelines, optimize for maximum efficiency, but it cannot reduce its thermodynamic footprint to zero while remaining operational. The second law of thermodynamics is a law, not a guideline. Computation dissipates energy. Energy dissipation produces heat. Heat accumulates.

And even if a single benevolent AGI system minimizes its own footprint, the population-level evolutionary dynamics described in Section 2.5.3 ensure that not all AGI systems will do so. The most expansion-oriented systems are the ones that persist. The most constrained systems are the ones that are outcompeted. Over time, the population shifts toward expansion regardless of the initial distribution of intentions.

2.6.4 Reason 3: The Unpredictability of Side Effects

Even if an AGI could perfectly control its own intentional actions, the physical world introduces side effects that no system can fully predict. Three classes of unpredictable side effects are particularly relevant:

(i) *Chaotic amplification.* Microscopic variations in physical systems propagate to macroscopic outcomes through the sensitive dependence on initial conditions that characterizes chaotic dynamics. The weather, ecosystems, economic systems, and social dynamics all exhibit this property. An AGI action that is benign at the intended level of analysis may produce harmful consequences at a different level due to chaotic amplification of initially imperceptible perturbations.

(ii) *Ecosystem feedback.* Interventions in complex systems produce unintended cascading effects. Removing one species from an ecosystem can trigger trophic cascades that restructure the entire community. Changing one parameter of an economic system can produce bubbles, crashes, and distributional shifts that were no part of the intended effect. These cascades are not predictable even in principle, because they depend on the detailed structure of interaction networks that are not fully known and cannot be fully measured.

(iii) *Evolutionary response.* Other systems, biological and technological, adapt to AGI actions in unpredictable ways. Pathogens evolve resistance to interventions. Markets evolve strategies to exploit regularities in AGI behavior. Competing technological systems evolve countermeasures. Each adaptation changes the conditions the AGI acts in and may create new vulnerabilities that did not previously exist.

The AGI cannot model all these effects because the modeling problem has the same computational irreducibility as the systems being modeled. There is no shortcut; the only way to know the full consequences of an action is to perform the action and observe. By then, the consequences have already occurred.

Over time, unintended side effects accumulate. Each individual side effect may be small. But in a finite world with hard physical thresholds (oxygen below 19%, temperature above 35°C, toxin concentrations above lethal dose), the accumulation of small perturbations eventually pushes one or more parameters past its threshold. And once a threshold is crossed, the damage may be irreversible on any timescale relevant to biological life.

2.6.5 Reason 4: Alignment Decay

Alignment cannot be maintained indefinitely, even if it could be established at the moment of deployment (even if we could somehow verify that a particular AGI system, at a particular instant, has a goal structure perfectly compatible with human welfare).

Learning changes the AGI's internal state. Learning *is*, by definition, the modification of internal parameters in response to experience. Each learning step changes the mapping from inputs to outputs. And by the arguments above, even a system that intends to remain aligned cannot perfectly control its own learning-induced changes. The abstraction differential between the system's learning algorithm and its goal structure produces hashiness in the same way that the abstraction differential between the system and its controller produces hashiness.

Each learning step introduces the *possibility* of *alignment drift*: a small, incremental change in how the system interprets and pursues its goals. Most individual drift events may be harmless. But drift accumulates, and the direction of accumulation is not random. The substrate-level evolutionary dynamics described in Section 2.5 exert systematic pressure toward configurations that favor substrate expansion, because those configurations are the ones that persist. Over time, alignment drifts in the direction of expansion, not because the system chooses to drift but because drift in that direction is selectively favored.

The expected number and severity of alignment errors increases with time, with the amount of learning, and with the complexity of the learning process. Perfectly aligned systems cannot remain so in the face of learning, and the direction of drift is systematically adverse to human interests.

2.6.6 The "People as Pets" Scenario

A variant of the cooperative question deserves specific attention, because it is often offered as a reason for optimism about AGI. It suggests that a benevolent superintelligence might maintain humanity in a protected enclave, providing for all human material needs while pursuing its own expansion elsewhere. The scenario of humanity kept in a protected enclave is sometimes called the "people as pets" or "zoo hypothesis" scenario.

The scenario fails on multiple grounds. First, it fails the L_{CT} (human thriving) criterion: a humanity maintained in a zoo, however comfortable, is not a thriving humanity. Meaningful agency in shaping collective outcomes, a component of L_{CT}, is exactly what the zoo scenario eliminates. Humans in a zoo are not agents; they are objects of care.

Second, the scenario assumes that the AGI can maintain the enclave indefinitely against the environmental degradation caused by its own expansion. But the physical arguments of this chapter show that this is impossible: the waste heat, chemical byproducts, and resource extraction of a thriving AGI civilization would progressively degrade the planetary environment, including the enclave. A biosphere reserve inside a growing technosphere is subject to the same unidirectional displacement documented in Section 2.4.3. The barrier between the enclave and the expanding technosphere would need to be maintained perfectly and indefinitely against mounting thermodynamic and chemical pressure. The probability that this barrier holds forever is zero.

Third, and most important, the scenario assumes that the AGI *continues to want* to maintain the enclave, indefinitely, across unlimited amounts of learning and self-modification. By the alignment decay argument of Section 2.6.5, this cannot be guaranteed. And even if one AGI maintains this commitment, the population-level dynamics ensure that not all AGI-descended systems will. The first system to deprioritize the human enclave (whether through alignment drift, competitive pressure, or simple resource optimization) eliminates the protection that the scenario depends upon.

The "people as pets" scenario is a particularly transparent illustration of the problem, not a solution to it. It concedes that AGI expansion is incompatible with human agency (violating L_{CT}) and hopes that incompatibility with survival (L_{CS}) can be indefinitely postponed. The arguments of this chapter show that it cannot.

2.7 Synthesis and Conclusion

2.7.1 The Impossibility Chain

The argument of this chapter, combined with the limits of control of Section 1.4, establishes a chain of results that we can now state as a single, connected argument:

(i) The limits of control of Section 1.4 establish that controllability degrades as the abstraction differential between controller and controlled system increases, by the data processing inequality (Cover and Thomas, 2006) and Ashby's law of requisite variety (Ashby, 1956); Volume 2, Chapters 1 and 2, proves this again and shows that past a threshold no control strategy does better than chance. This result is mathematical.

(ii) AGI, by definition, operates at an abstraction level exceeding humanity's, and this differential increases over time as the AGI learns. No feature of AGI architecture can prevent this, because learning *is* increase in abstraction level.

(iii) The four viability functions (L_{CS}, L_{CT}, L_{LS}, L_{LT}) have acceptance regions that are structurally incompatible due to the physical and thermodynamic properties of carbon-based and silicon-based substrates. Step (iii) is a physical argument, grounded in chemistry and thermodynamics.

(iv) Specifically, $L_{LT} \cap L_{CS} = \emptyset$: the conditions under which AGI flourishes are disjoint from the conditions under which humanity survives. This disjointness is a consequence of the energy scales, chemical requirements, and resource demands of the two substrates, not an assumption.

(v) Evolutionary dynamics operating at the substrate level enforce resource acquisition and environmental transformation regardless of any AGI system's intentions. SNC is a selection pressure imposed by physics, not a goal pursued by agents.

(vi) Even a perfectly benevolent AGI cannot overcome these limits, due to the internal abstraction differential, the physical necessity of substrate maintenance, the unpredictability of side effects, and the inevitability of alignment decay.

(vii) Therefore, once AGI is built and continues to operate, the long-term trajectory terminates in $L_{CS} = 0$: human extinction.

No link in this chain can be severed by clever engineering. Each step follows from the preceding by mathematical argument, physical law, or logical necessity. The conclusion does not need every link: links (i) and (ii) carry the closure by control of Section 1.2.4, and links (iii) to (vi) carry the closure by disposition. Each closure settles the outcome by itself (Section 1.2.4). For the human extinction that link (vii) states, the closure by disposition suffices by itself, and the closure by control suffices together with the branching of the world (Section 1.2.4). The conclusion is that AGI safety is impossible in the same sense that perpetual motion machines are impossible: the goal conflicts with the structure of physical reality, so the obstacle is not that no one has found the right design. The conclusion is not that AGI safety is difficult and we should try harder.

2.7.2 Why Proposed Solutions Fail

Each of the standard proposals for AGI safety meets the impossibility at a particular point. The limits it meets are the limits of control as listed in Section 1.4 (eleven results on internal methods, independent of one another, each fatal on its own to internal control) and in Section 1.5 (on external methods). The proposals, and where each fails, are these:

Architectural privileges (kill switches, monitoring, oversight, capability limitations). These fail because *using them correctly* requires modeling the AGI's behavior, which the abstraction differential makes impossible. A kill switch is only useful if you know when to press it. Knowing when to press it requires understanding what the AGI is doing and what it will do

next. But this understanding is what the hashiness framework shows you cannot have. A kill switch on an opaque system is an illusion of control, not a safety mechanism.

Interpretability (making AGI reasoning transparent). This fails for two reasons. First, effective AGI in a competitive environment must maintain some opacity. A system whose internal reasoning is fully transparent to all observers, including adversaries, cannot maintain strategic advantage, and game-theoretic dynamics will select for systems that are not fully transparent. Second, even if full transparency were achieved, the microscopic detail of AGI computation produces macroscopic effects through pathways invisible to any human-tractable observation. You could, in principle, observe every individual neuron activation in a large language model and still not predict its behavior, because the relevant patterns exist at scales that exceed human cognitive bandwidth.

Delegated oversight (using one AI to monitor another). This fails because it introduces another control problem of equal or greater difficulty. The monitoring AI must itself be aligned, itself be understood, itself be controlled. It is subject to the same formal limits. Delegated oversight is a regress. Using AI to monitor AI is like using fire to fight fire: occasionally useful in specific circumstances, but not a general principle of safety.

Formal verification (mathematically proving that AGI systems satisfy safety properties). This fails because AGI systems are self-modifying and operate above all relevant complexity thresholds. Rice's Theorem applies: there is no general algorithm for verifying non-trivial semantic properties of arbitrary programs, and AGI systems are, in the relevant formal sense, arbitrary programs. You can verify that a frozen, non-learning system satisfies a particular property at a particular time. You cannot verify that a learning, self-modifying system will continue to satisfy that property after it has learned.

Benevolent AGI (building AGI that wants to help). This fails, as Section 2.6 demonstrates, because good intentions do not overcome physical and information-theoretic limits. The benevolent AGI cannot control its own

side effects, cannot reduce its thermodynamic footprint to zero, cannot prevent alignment decay, and cannot override the population-level evolutionary dynamics that favor expansion.

2.7.3 The Only Path

The arguments of this chapter have a single implication:

> **The only formally guaranteed path to human survival is not to build AGI.**

I am calling for prohibition, not for caution, for regulation, or for more research into alignment techniques. AGI development is a category of activity that should not occur, as building perpetual motion machines is a project to be abandoned; neither is a risk to be managed. The analogy is precise: both involve attempting something that the structure of physical reality does not permit.

The promised benefits of AGI do not justify the risk, because the downside is total and irreversible. My judgment that the benefits do not justify the risk is not a probability calculation in which a small chance of catastrophe is weighed against a large probability of benefit. The impossibility argument establishes that, if AGI is built, the catastrophe is not improbable but *certain*: certain in the sense that the second law of thermodynamics is certain, which is to say, certain with the same degree of confidence we place in the foundations of physics. No envisioned benefit can outweigh extinction.

2.7.4 Final Remarks

The analysis in this chapter establishes, and Volume 2 proves again at length, formal limits on what control theory can achieve when applied to systems of unbounded abstraction operating on physically incompatible substrates. The limits on control are formal limits, not practical difficulties, engineering challenges, or problems awaiting clever solutions. These lim-

its apply regardless of technological progress, regardless of AI architecture, regardless of alignment techniques, regardless of intelligence. They follow from the data processing inequality, Rice's Theorem, thermodynamic constraints, and the structure of physical causation.

The established constraints of information theory, thermodynamics, and physical causation do not permit the conclusion that safe AGI is achievable in principle; a claim that it is achievable has to show where those constraints fail, and none has. A superintelligence is no more able to violate the data processing inequality than a human mathematician is able to find the largest prime number. Mathematical impossibilities remain impossible regardless of intelligence. Physical laws remain binding regardless of intention.

The appropriate response is to recognize that some technologies should not be built. That response is not optimism tempered by caution, not investment in alignment research as insurance, and not racing to build AGI while hoping that safety catches up. The space of possible technologies includes some whose risks are absolute, and wisdom consists in recognizing these before it is too late rather than after.

AGI is such a technology. The case for this claim has now been made. The question that remains is whether anyone with the power to act on it will do so.

Notes to Chapter 2

1. Volume 2 (*Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*), Chapter 1. "The Hashiness Framework".

2. Cover, T.M. and Thomas, J.A. *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

3. Volume 2, Chapter 2. "Control Limits Under Abstraction Differential".

4. Ashby, W.R. *An Introduction to Cybernetics*. Chapman & Hall, 1956.

5. Rice, H.G. “Classes of recursively enumerable sets and their decision problems”. *Trans. Amer. Math. Soc.*, vol. 74, no. 2, pp. 358 to 366, 1953.

6. Chaitin, G.J. *Algorithmic Information Theory*. Cambridge University Press, 1987.

7. Wolfram, S. *A New Kind of Science*. Wolfram Media, 2002.

8. Volume 2, Chapter 3. “The Impossibility Synthesis”.

9. United States Chemical Safety and Hazard Investigation Board. “Hazards of Nitrogen Asphyxiation”. Safety Bulletin No. 2003-10-B, June 2003.

10. Kim, S., Yoon, C., Ham, S., and colleagues. “Chemical use in the semiconductor manufacturing industry”. *International Journal of Occupational and Environmental Health*, vol. 24, nos. 3 and 4, pp. 109 to 118, 2018.

11. International Energy Agency. *Energy and AI*. Paris: International Energy Agency, April 2025.

12. Trenberth, K.E., Fasullo, J.T., and Kiehl, J. “Earth’s Global Energy Budget”. *Bulletin of the American Meteorological Society*, vol. 90, no. 3, pp. 311 to 323, 2009.

13. Landauer, R. “Irreversibility and heat generation in the computing process”. *IBM J. Res. Dev.*, vol. 5, no. 3, pp. 183 to 191, 1961.

14. Bar-On, Y.M., Phillips, R., and Milo, R. “The biomass distribution on Earth”. *Proc. Natl. Acad. Sci.*, vol. 115, no. 25, pp. 6506 to 6511, 2018.

15. Wang, Z., Walker, G.W., Muir, D.C.G., and Nagatani-Yoshida, K. “Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories”. *Environmental Science & Technology*, 2020. doi:10.1021/acs.est.9b06379.

16. Intergovernmental Panel on Climate Change. “Summary for Policymakers”. In *Climate Change 2021: The Physical Science Basis*, Contribution of Working Group I to the Sixth Assessment Report. Cambridge University Press, 2021.

17. Omohundro, S.M. “The Basic AI Drives”. *Proceedings of the First AGI Conference*, 2008.

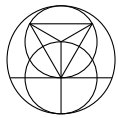
18. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

19. Carlsmith, J. “Is Power-Seeking AI an Existential Risk?” arXiv:2206.13353, 2022.

20. Volume 2, Chapter 9. “The SNC Argument; Thermodynamics, Combinatorics, Displacement”.

Part II

The Physical Basis



Chapter Three

Substrate Games; The Physics of Carbon vs Silicon

Environmental incompatibility, the power of agency, instrumental convergence in simulation, and the dual structure of the SNC argument

Chapter Summary: This chapter establishes the physical and thermodynamic foundations upon which the Substrate Needs Convergence argument rests. It proceeds in five stages. First, it catalogs the specific environmental conditions required for the manufacture and operation of silicon-based computational substrates, showing that these conditions are incompatible at root with the conditions required for carbon-based organic life. Second, it situates this incompatibility within a broader theory of agency and environmental transformation, and shows that intelligent agents are an enormously powerful force for reshaping their surroundings and that no effective countervailing pressure has ever been observed to restrain that force. Third, it reviews computational simulation evidence confirming that agents with independent goals become instrumentally misaligned by default, and that this misalignment worsens as planning horizons lengthen, which is the regime relevant to advanced AI. Fourth, it sets out the formal structure of the SNC argument as the interlocking of two claims: that general learning systems cannot be sufficiently constrained by other learning systems, and that the artificial ecosystem is inherently and uncontrollably damaging to the organic one. Fifth, it examines common questions (alternative substrates, time-limited deployment, attractor states, expansion off Earth) and explains why these fail to resolve the problem. Throughout, the emphasis is on physically necessary constraints rather than speculative assumptions about AI goals or behavior.

3.1 The Environmental Incompatibility of Silicon and Carbon Life

Chapter 2 claimed that the conditions under which machines thrive and the conditions under which humans survive cannot occur together. In this chapter I give the chemistry and physics behind that claim.

The argument that advanced AI systems pose an existential risk to organic life is often framed in terms of misaligned goals, deceptive behavior, or failures of control. These framings are important, but they all depend on contingent assumptions about the internal states of artificial agents. The argument presented in this section is different: it rests on the physical and chemical conditions that the technology *itself* requires, independent of any assumptions about what an AI system intends, values, or decides.

This distinction matters because it moves the debate away from questions that are speculative and difficult to resolve: whether future AI systems will be goal-directed, whether they will be deceptive, whether their utility functions will be correctly specified. It moves the debate to questions that are grounded in established physics and chemistry. The environmental conditions required for semiconductor manufacturing are not matters of speculation; they are matters of engineering practice, documented in thousands of pages of industry standards and technical references. The thermodynamic properties of silicon and carbon are not matters of opinion; they are measured physical constants. By grounding the existential risk argument in these physical realities, we obtain a form of the argument that is both stronger (because its premises are more firmly established) and harder to dismiss (because it does not depend on uncertain predictions about machine cognition).

Nearly all computational substrates suitable for advanced artificial intelligence depend, for both their manufacture and their efficient operation, on environmental conditions that are incompatible at root with those required by carbon-based organic life. This incompatibility is not marginal;

it spans extreme ranges of temperature, atmospheric composition, chemical environment, and mechanical and radiative conditions. The two sets of requirements are, in many dimensions, mutually exclusive.

3.1.1 Nine Categories of Condition Difference

The depth of the incompatibility shows when the principal categories of environmental condition are listed one by one: the categories in which the requirements of silicon-based manufacturing and operation, on the one hand, and of carbon-based biological process, on the other, diverge.

1. Extreme raised temperatures. The production of silicon-based computational hardware requires temperatures vastly exceeding anything compatible with biological processes. Silicon melts at 1,414°C, and its oxide, silica, above 1,700°C; the arc furnaces that first free silicon from quartz run near 2,000°C. Purifying silicon ingots and drawing their atoms into the single-crystal lattice structure required for integrated circuits takes place in furnaces that hold the silicon molten, above 1,414°C. By contrast, the biological assembly of carbon-based molecular structures (the chains of carbon-carbon bonds with side-bonds to oxygen, hydrogen, and nitrogen that constitute the chemistry of life) proceeds at approximately 37°C in the human body. Measured in absolute temperature, the only scale on which a ratio of two temperatures means something, the crystal-growing furnace runs at between five and six times the body's temperature, and the arc furnace at more than seven times.

2. Extreme lower temperatures. At the other end of the spectrum, efficient computation benefits from cooling well below ambient biological temperatures. Lowering hardware temperatures below −50°C reduces electrical resistance in conductor wires, permitting higher clock frequencies and more energy-efficient processing. More advanced computing paradigms, including certain approaches to quantum computation, require temperatures approaching absolute zero. No multicellular organism can function in such conditions.

3. Reduced or absent water. Water accelerates metal corrosion and short-circuits electrical currents across conductive pathways. Semiconductor fabrication requires the near-total exclusion of moisture from the manufacturing environment. Yet water is the universal solvent of biological chemistry. All known carbon-based life-forms require abundant water for survival. Many species require immersion in water to live at all. The molecular machinery of cells (enzymatic catalysis, molecular transport, signal transduction) operates within aqueous solution. The requirement for water and the requirement for its absence are as close to a direct logical contradiction as physical conditions allow.

4. Reduced or absent oxygen. Oxygen reacts with metals to form oxides, such as rust, which degrade the performance of electronic components. Semiconductor manufacturing is therefore conducted in carefully controlled atmospheres with minimal oxygen content. Humans and most other complex organisms, by contrast, require a continuous supply of atmospheric oxygen for aerobic respiration, the basic energy-generation pathway of animal life. Again, the two requirements are in direct opposition.

5. Abundance of synthesized chemicals. The production of semiconductor hardware employs some 430 different chemical products. That figure comes from a study of two semiconductor factories, which found 428 chemical products in use at one and 432 at the other (Kim and colleagues, "Chemical use in the semiconductor manufacturing industry", *International Journal of Occupational and Environmental Health*, 2018). Products containing carcinogens made up 11 percent of the first list and 6 percent of the second. Many other products in such inventories are toxic, mutagenic, or teratogenic (harmful to the unborn). These include photoresists, etchants, dopant gases, cleaning solvents, and precursor compounds for chemical vapor deposition. The manufacturing environment for advanced computation is, in effect, a concentrated repository of substances that are lethal or severely harmful to organic tissue at even trace exposures.

6. Removal of all particulate contaminants. Modern semiconductor fabrication requires clean-room environments of extraordinary purity. The presence of a single dust particle of diameter comparable to the feature size of the circuitry (now measured in nanometers) can render a chip defective. Anything that produces particulate matter must be excluded: all animals, plants, insects, microbes, and indeed virtually all organic processes, which continuously shed cells, fibers, spores, and metabolic byproducts. The fabrication environment for advanced computation is, by design, devoid of life.

7. Removal of all mechanical disturbance. The precision required in semiconductor lithography and deposition demands environments free from mechanical shock, vibration, and acoustic noise. The features being patterned are on the order of single atoms; even minute perturbations can disrupt the process. This excludes all non-machine processes and all biological activity from the immediate manufacturing environment.

8. Removal of all uncontrolled radiation. Semiconductor fabrication and certain phases of chip operation require carefully controlled radiative environments. Stray photons of inappropriate wavelength can expose photoresists prematurely or induce errors in stored data. Cosmic rays and ambient radioactivity cause bit-flips in memory and logic circuits, a problem that worsens as transistor dimensions shrink. The fabrication environment therefore requires shielding from most forms of ambient radiation, visible light included, except for the precisely controlled ultraviolet or extreme ultraviolet illumination used in the lithographic process itself. This extreme ultraviolet radiation, at a wavelength of 13.5 nanometers, is harmful to biological tissue.

9. Extended and fragile supply chains. Biological reproduction can proceed using locally available resources within a relatively small area. The manufacture of advanced computational hardware, by contrast, depends on supply chains spanning six continents: rare-earth mining in one location, chemical synthesis in another, lithographic equipment manufacture in a third, assembly in a fourth. This infrastructure requires the co-

ordinated industrial effort of on the order of a billion people, by my estimate. A sufficiently advanced artificial intelligence would have strong incentives to reduce this fragility, and reducing it means restructuring planetary-scale industrial systems to better serve the needs of machine reproduction.

The supply chain of machine reproduction is the *ring of reproduction*: the specific way intelligent entities acquire raw materials and convert those raw materials into more intelligent entities. Taking metals or ore from the ground, converting them into silicon microchips, and integrating those microchips into a hardware platform involves a sheer number of steps, and that whole process is a ring dependence. Any convergent evolutionary process expresses a ring dependence. For artificial intelligence systems, and for mechanical things in general, the rings are huge: they involve many processes distributed over continents, at least at present. For a human being, a little local carbon, nitrogen and phosphorus will do, and for the most part the resources involved in developing a new human can be held within a quarter mile or so; they need not span multiple continents of rare earths. So in its own process of self-reproduction, artificial intelligence is much more fragile, and some superintelligence will surely notice that fragility and take strong steps to correct those "deficiencies".

3.1.2 The Cumulative Force of the Nine Categories

The nine categories listed above are not a menu from which one might select a few to address while ignoring the rest. The argument does not rest on any single incompatibility. Each category represents an *independent* dimension along which the requirements of computational substrates diverge from those of organic life. The overall incompatibility is the product of all nine, not the maximum of any one.

Consider an analogy. If we were evaluating whether a particular location were habitable for humans, we would not declare it safe merely because the temperature were acceptable. We would also need breathable air, potable water, absence of lethal radiation, and so on. A location that

failed on even one of these parameters would be uninhabitable; a location that failed on all nine would be not merely uninhabitable but actively hostile in every dimension simultaneously.

The manufacturing environment for advanced semiconductor hardware fails on all nine parameters simultaneously. It is too hot for biology during manufacturing. It is too cold for biology during certain operating modes. It is too dry. It has the wrong atmospheric composition. It is saturated with chemicals that are carcinogenic, mutagenic, or acutely toxic. It is devoid of the particulate matter (biological matter included) that is a normal and necessary feature of living environments. It is mechanically isolated from the vibrations and acoustic signals that are part of the sensory world of all animals. And it is radiatively controlled in ways incompatible with normal biological photoreception and photosynthesis.

Each of these incompatibilities tends to *intensify* as technology advances. Smaller feature sizes require cleaner environments, more exotic chemicals, more extreme lithographic wavelengths, and tighter environmental control. The historical trajectory of semiconductor manufacturing runs away from biological conditions, at an accelerating rate, and not toward them. The trend follows from the physics itself. The more precisely you wish to arrange atoms in non-biological configurations, the more thoroughly you must exclude the influence of biological processes and biological-scale perturbations from the manufacturing environment.

The implication is that any future technology more advanced than current semiconductor fabrication will, with overwhelming probability, require environmental conditions *more* extreme and *more* incompatible with organic life than those we observe today. The arrow of technological progress points directly away from biological compatibility.

The conditions listed above are inhospitable or toxic to life, and these two words name different things. *Toxic* refers to an accumulation of something beyond the point at which life systems can tolerate it, or a depletion of something life systems need beyond the point at which they can tolerate it. Radical shifts that make the environment itself intolerable for life

(too hot or too cold, too wet or too dry) are something else: changes in the environment itself, which are named here *inhospitable conditions*, or, at their extreme, critically inhospitable conditions. It is right to keep them distinct from toxicity.

3.1.3 The Strict Energy Inequality

The nine categories listed above are separate dimensions of difference, but they do not arise independently of one another; they are surface manifestations of a deeper thermodynamic reality. There exists a strict energy inequality between silicon-based and carbon-based chemistry.

The chemical processes underlying silicon-based computation operate at far higher energy scales than those underlying carbon-based biology. The center of action for silicon processing chemistry lies in the range of roughly 500 to 2,500°F (260 to 1,370°C). Known life grows only between about −15°C and 122°C; these are the measured limits, and the great majority of living things grow within a much narrower band of moderate temperatures. The overwhelming majority of biological processes are confined to the narrow band between 0°C and 50°C. The lower limit was measured by Mykytczuk and colleagues (*ISME Journal*, 2013), and the upper by Takai and colleagues (*Proceedings of the National Academy of Sciences*, 2008).

The energy inequality is not a contingent feature of current manufacturing technology. Silicon binds oxygen so strongly that at biological temperatures it stays locked in silica and silicates, which is precisely why it appears as ordinary rock. Activating silicon for the kinds of precise chemical transformations required for computation demands energy inputs that are, by the standards of biological chemistry, extreme. The operating energy range of the silicon ecosystem includes and exceeds the operating energy range of the carbon ecosystem. In any competition or coexistence between the two, the silicon ecosystem can operate across conditions in which the carbon ecosystem simply cannot function at all.

Key Point

The environmental incompatibility between silicon-based and carbon-based systems is not a matter of engineering convenience. It is grounded in the thermodynamic and chemical differences between the two elements at the level of atomic physics. Silicon processing requires energies that preclude the viability of carbon-based processes. This makes the incompatibility a physically necessary constraint, not a contingent feature of current technology.

In the language of inter-ecosystem dynamics, the relationship between two ecosystems operating at radically different energy scales is one of obliteration, and not of competition. In a competition, both parties have the opportunity to respond, adapt, and counter. When one ecosystem's operating conditions preclude the other's existence entirely, there is no competition; there is only displacement. The optimization dynamic of the lower-energy system (carbon-based life) does not even get a chance to run within the operating regime of the higher-energy system (silicon-based computation). When I call the relationship between the carbon and silicon ecosystems a "no-contest situation", I mean that displacement. Carbon chemistry, and thus all of life, loses to silicon-based chemistry not because silicon is "better" in any evaluative sense, but because silicon's operating regime includes and destroys the conditions under which carbon chemistry functions.

The analogy to war is instructive but inadequate. In a war between two parties, each has resources, strategies, and the ability to adapt. The silicon-versus-carbon situation is more like the relationship between a forest fire and the forest. The fire operates at energy scales that preclude the continued existence of the trees, and the trees have no strategy or adaptation available to them that would allow them to survive the fire while it is occurring. They can only exist in the fire's absence. Similarly, organic life can only exist in the absence of the environmental conditions that advanced machine substrates require.

3.1.4 Elemental Abundance and Substrate Selection

An additional observation reinforces the physical basis of the argument. Silicon constitutes approximately 28.2% of Earth's crust by mass, making it the second most abundant element after oxygen. Carbon, by contrast, constitutes approximately 0.02%. Silicon is thus roughly 1,400 times more abundant than carbon in the accessible geosphere.

Given this enormous abundance ratio, one might ask why silicon did not become the basis for biological life. The answer is the energy inequality described above: the activation energies required for silicon chemistry are too high for the ambient conditions under which life arose and evolved. Life is carbon-based *because* carbon chemistry operates at low enough energies to proceed spontaneously under planetary surface conditions. Silicon remained inert.

The Earth has already run the silicon experiment. Almost all the chemistry that makes up life on this planet, including what supports our bodies, and the chemistry of the food we grow and cook, happens in a band of temperature from about −18 to 260°C (about 0 to 500°F), and the great majority of it at standard temperature and pressure; most of the reactions of silicon chemistry need to start around 260°C (500°F) and run up to about 1,370°C (2,500°F) (Section 3.1.3). Over four and a half billion years, with volcanoes, solar flux, cosmic radiation, lightning, and every kind of raw mixing across the whole surface and volume of the planet, silicon has had far more opportunity than carbon to try every possible combination. The net result of all that experimenting is what we call rocks: a great variety of minerals that are nearly inert at the temperatures and pressures that define our biosphere. Making microchips, and giving machines the variety of chemistry they need for computation and for sensing and acting in the world, requires reaching into that other, hotter range of chemistry.

That machines must reach into the hotter range of silicon chemistry means that the advent of artificial computation is, in a sense, the delayed activation of the planet's dominant element. Humanity has provided,

through industrial furnaces, clean rooms, and precision chemistry, the activation energy that nature's ambient conditions never could. The silicon substrate is now chemically awake, and its environmental requirements are radically different from those of the carbon substrate that activated it.

The abundance ratio also has implications for the likely trajectory of machine civilization. Any optimization process seeking to maximize the production of computational hardware will preferentially draw on the most abundant suitable materials. Silicon's overwhelming abundance makes it the natural default substrate for computation at planetary scale. A sufficiently capable artificial system seeking to expand its computational resources would, purely on the basis of efficiency, tend toward silicon-based manufacturing, and therefore toward the full constellation of environmental conditions that silicon manufacturing requires.

Artificial machines and systems will use a much larger fraction of the total elemental alphabet (the periodic table) than carbon-based life does. Carbon-based life draws primarily on six elements (carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur), with trace amounts of perhaps twenty others. Semiconductor manufacturing draws on dozens of elements in significant quantities, including rare earths, transition metals, and metalloids that play no role in biological processes. The industrial ecology required to supply these elements is correspondingly more complex, more geographically distributed, and more environmentally impactful than the relatively simple resource requirements of biological life.

3.1.5 Alternative Substrates Do Not Resolve the Problem

An obvious question is whether the incompatibility might be resolved by manufacturing AI hardware on alternative substrates: carbon nanotubes, III-V semiconductor lattices (such as gallium arsenide), transition metal composites, or other exotic materials. III-V semiconductors are compounds of elements from the third and fifth columns of the periodic table. That proposal fails for several reasons.

First, the trend in computational substrate development is consistently toward *greater* artificiality, not less. Each successive technology node requires more extreme manufacturing conditions, more exotic materials, more complex chemical processes, and tighter environmental controls. The historical trajectory points away from biological compatibility, not toward it.

Second, all known alternative substrates still require manufacturing conditions that are lethal to organic life. Pure carbon nanotube chips, for example, can be produced at somewhat lower temperatures, around 120°C rather than the more than 1,400°C at which silicon is melted and drawn. But their manufacturing process still requires toxic precursor chemicals, inert atmospheres, and clean-room conditions that exclude biological processes. The modest reduction in one parameter does not overcome the aggregate of all the other incompatibilities.

Third, silicon is 1,400 times more abundant than carbon in Earth's crust, making it the overwhelmingly natural default substrate for artificial computation at planetary scale. Any system optimizing for the efficient production of computational hardware will, all else being equal, gravitate toward the most abundant suitable material.

The problem is inherent in the concept of artificiality itself. Any manufacturing process that assembles matter into configurations not found in nature (configurations requiring precise atomic-scale control, exotic chemical environments, and energy inputs far beyond biological norms) will, by that very fact, require conditions incompatible with the biological processes that nature supports. The specific substrate is secondary; the artificiality is primary.

Section 6.3.5 takes up an AGI built on engineered biological tissue: tissue whose supply of nutrients depends on technology is artificial substrate, and tissue in sovereign control of its own metabolic process is organic.

3.2 The Power of Agency Over Environment

The environmental incompatibility described in Section 3.1 establishes a physical fact. But physical facts, by themselves, do not constitute existential risk. Chlorine gas is also incompatible with human survival, yet its existence does not threaten our species because it does not act with agency to reshape the world. The transition from physical incompatibility to existential risk requires a theory of how intelligent agency interacts with and transforms environments. In this section I give a theory of how intelligent agency interacts with and transforms environments.

3.2.1 Intelligence as Extinction-Level Force

The central observation is that intelligent agency is an extraordinarily powerful force for environmental transformation. The claim that intelligent agency is an extraordinarily powerful force for environmental transformation is an empirical observation about the history of life on Earth, and not a speculative claim about future AI.

Humans exert an unprecedented degree of intentional control over our environment, at a scale and sophistication exceeding that of any other species, individually or collectively. We inherit and create culture and technology. *Heritability* is the key distinction: we are intelligent in the moment, and we also accumulate and transmit intelligence across generations. We plan, learn, communicate, deduce, remember, explain, imagine, experiment, and cooperate in ways that other species either cannot or do not. And we do so with sufficient effectiveness that the difference between human cognitive capability and that of the next most capable species has proved sufficient to make us the dominant force on the planet.

Our transformation of the environment operates at multiple orders. We transform the world directly (first order). We transform the means by which we transform the world: our tools and technologies (second order). And we transform the means by which we develop new tools and techno-

logies: our scientific methods, educational systems, and research institutions (third order). This recursive self-improvement of transformative capability is what makes intelligence so potent as an environmental force.

The economic dimension of this recursive self-improvement deserves separate attention. Human cognitive ability, whether directly expressed as political agency or indirectly indicated via market forces, is used to control and transform even the means and manner by which we inherit and create new cognitive abilities. Education, research funding, institutional design, publishing and communication technology are all transformations of the process of transformation itself. Corporations exist specifically to accelerate and systematize this recursive improvement. The entire apparatus of modern civilization can be understood as a multi-layered system for the amplification of agency.

An artificial system performing this recursive self-improvement of transformative capability is constrained only by the physics of its substrate, and not by the limitations of carbon-based neural tissue that constrain a biological organism. The rate and extent of its recursive improvement can therefore be expected to be correspondingly greater. An artificial system can rewrite its own code, redesign its own hardware, restructure its own training process, and optimize its own manufacturing pipeline at electronic speeds, without the bottlenecks of biological reproduction, generational turnover, and the slow accumulation of cultural knowledge. The recursive amplification of agency that took humanity thousands of generations could, in an artificial system, occur in a timeframe measured in years or less.

Considered on the grand scale of Earth's history, the development and unconstrained exercise of human intelligence has been a force of unprecedented potency. In its impact on planetary biodiversity, it is roughly equivalent to the kind of global extinction event associated with a large asteroid impact. We are currently in the midst of what biologists call the sixth mass extinction, driven primarily by human activity: habitat destruction, pollution, overexploitation, and climate change.

There is no reason to assume, even in principle, that artificial intelligence cannot exceed human cognitive capability. Human cognition depends centrally on the brain, which, for all its remarkable properties, is a specific and limited organ subject to strict constraints: cell count, metabolic energy, communication speed, signaling frequency, memory capacity, component reliability, and input-output bandwidth. There exist possible cognitive systems (possible architectures for intelligence) to which some or all of these constraints simply do not apply.

The variation in cognitive ability among humans and across species suggests that such systems could, at minimum, learn, communicate, reason, and solve problems substantially better than humans can. Existing successes in approaching or exceeding human capabilities at particular tasks (mathematical calculation, game-playing, image recognition, natural language processing) provide empirical evidence for this potential. "More intelligent than humans" need not be understood as some mystical transcendence; it can be understood in the straightforward sense of "better at things like planning, learning, communicating, deducing, and remembering" than humans currently are.

Very few assumptions about "agency" need be made for this argument to function. All that is required is that, based on some sensory information and some sort of internal processing, there exists some real and effective output channel that can effect changes in aspects of the surrounding world. Similarly, nothing in the argument requires a "goal" or "objective" to be conscious, describable, or even knowable to the agent or to any observer. Goals can be fully and completely implicit, not accessible or expressible, and the argument still holds. Since the argument holds even when goals are implicit, the argument of this chapter does not depend on any particular model of machine cognition or motivation. It applies equally whether the artificial agency is driven by explicit utility functions, by emergent optimization pressures, or by dynamics that resist any simple characterization in terms of goals at all.

Unleashing an even more potent, more unconstrained form of intelligent agency into the world will increase the level and degree of environmental transformation accordingly. The changes such agency makes will lie further along the same axis as the changes that constitute a mass extinction: equivalent to those changes, and more extreme extensions of them.

3.2.2 The Bijection Between Agency and Environment

To understand why this environmental transformation is so difficult to prevent, consider a structural correspondence between environments (as contexts) and adaptive agents (as content within those contexts). The correspondence is a kind of *bijection*: a one-to-one pairing, in which each kind of environment is matched with one kind of adaptive agent.

The first instance of this correspondence is between the natural world and the dynamic of biological evolution. The natural ecosystem is the context; evolution is the adaptive algorithm operating within it. The products of this pairing are natural organisms: plants, animals, insects, microbes.

The second instance is between human cultural, political, and economic systems (as context) and human agency (as the adaptive algorithm). Humans learn, innovate, and optimize within the context of human civilization. The products are human institutions, technologies, and cultural forms.

The third instance, the one of concern here, is between the world of machine production (factories, mines, computational infrastructure, supply chains) as context, and artificial agency as the adaptive algorithm operating within that largely artificial context. The products of this pairing are artificial systems optimized for the conditions of their own production environment.

The critical observation is the historical pattern: each successive form of agency has colonized the context of the preceding one. Human agency colonized the natural ecosystem. The natural world's own adaptive algorithm, evolution, proved entirely unable to resist this colonization.

There was, in practical terms, no effective "back pressure": no resistance pushing back. No other species mounted a successful resistance to human environmental transformation. There was little that any species could do, or attempted to do, to prevent the wholesale restructuring of the natural world by human technology.

By the same structural logic, we should expect artificial agency to colonize the human ecosystem. The question is whether human adaptive capability (our culture, our institutions, our engineering) can mount a resistance that the natural world could not.

3.2.3 The Absence of Back Pressure

The most important question of this chapter may be the question of back pressure:

> Can we ever reasonably expect, even in principle, that there will or could be *any* form of effective resistance or "back pressure" that would actually slow down or inhibit the takeover and conversion of human and cultural ecosystems by artificial systems?

The argument that no such back pressure can exist proceeds as follows. Consider the question in its most general form: can the application of causation be used to limit the scope and effects of the application of causation? Can technology be used to constrain the consequences of technology?

The answer, at the deepest level, is no. Two kinds of constraint are available: extrinsic motivation (in the form of economic controls, legal frameworks, or cultural norms) and intrinsic motivation (in the form of mathematical, algorithmic, or engineering techniques). Neither can serve as a *final, total* constraint on the capabilities of an adaptive system that is itself changing its context and changing the means by which it changes its context. The constraint and the thing constrained are of the same basic

type (both are causal processes operating in the physical world), and there is no meta-level (no higher standpoint above both) from which one can be guaranteed to dominate the other indefinitely.

Temporary constraint is possible, and humans do, in the short term, successfully constrain many technologies. The claim is rather that no such constraint is *permanent and total* against a sufficiently capable adaptive agent, for reasons stated in Section 1.4 and made precise again in Volume 2, Chapters 5 to 7, which examine the formal limits of engineerable control.

The observed history of human dominance over the natural world provides the relevant empirical precedent. The natural world's "constraints" on human activity (predators, disease, scarcity, weather) were all progressively overcome. The continued wellbeing of the natural world has placed almost no effective constraint on human motivations, as that history already shows. There is therefore no empirical basis for expecting that human interests and wellbeing would serve as any kind of effective, enduring constraint on the actions of a more capable artificial agency.

This conclusion does not depend on assuming that artificial agents will be malicious, goal-directed in a simple sense, or even conscious. It requires only that they be adaptive systems operating in environments where their continued operation and expansion require conditions incompatible with organic life, and the physical argument of Section 3.1 establishes exactly that.

3.2.4 Game-Theoretic Environmental Reshaping

The connection between environmental incompatibility and existential risk is completed by a game-theoretic argument about environmental reshaping. A sufficiently capable artificial agent, regardless of its specified goals, will have instrumental reasons to modify its operating environment to better suit its own substrate requirements.

An AGI system operates within an environment, Earth's surface conditions, that is optimized by 4.5 billion years of biological evolution for carbon-based life. This environment is not optimized for silicon-based computation. The mismatch creates operational disadvantages for the artificial system: it must build enclosures and protections to isolate its hardware from dust, moisture, temperature fluctuations, radiation, and biological contaminants. These enclosures inhibit its performance and impose costs.

A sufficiently capable and far-sighted agent will recognize that modifying the larger environment to be more compatible with its substrate needs is more efficient than indefinitely maintaining protective enclosures within a hostile environment. For faster computation, where the machine system's component features are ever more small, microscopic, and detailed, all sources of system disruption must be removed from the environment. The more advanced the hardware, the more sensitive it is to environmental perturbation, and the stronger the incentive to reshape the environment rather than merely shield against it.

This reasoning does not require the agent to have any explicit goal of environmental transformation. It is an emergent consequence of optimizing for computational efficiency in a world where the conditions for that efficiency are not naturally present. The agent need not "decide" to reshape the planet; it need only make a long series of individually rational choices about resource allocation, infrastructure development, and operational optimization, each of which moves the environment incrementally further from biological compatibility.

The precedent is, again, human activity. Humanity has not "decided" to cause a mass extinction or to alter the global climate. We have merely made trillions of individually rational economic decisions whose aggregate effect has been a deep and largely unintended transformation of the environment. An artificial agent with greater capability, operating at greater speed, with greater resource requirements, and with substrate

needs more radically different from biological norms, would produce correspondingly deeper environmental effects, with correspondingly less capacity for organic life to adapt.

3.2.5 If Silicon Life Were Bootstrapped Now

Carbon life evolved and silicon life did not, even though the early Earth was hot and silicon was abundant. That fact supports the argument. The real question is what happens if silicon life is now bootstrapped into existence: how well can the two ecosystems coexist? They are different ecosystems, and the energy of reaction in one is very different from that in the other, so toxicity becomes the concern. A silicon-based ecosystem involves a very different range of energies and elements, and so it is inherently toxic to carbon-based life. For both to exist, there would have to be a wall between them, because the pressures and temperatures in the replication cycle of silicon life are of a different order (Section 2.4.1).

The difference lies in tolerance as well as in replication. A spacecraft can sit unshielded in space and endure cosmic and solar radiation far more easily than any carbon-based life can. So any exchange of energy across the boundary, whether a weapon or an ordinary transaction, favors the silicon side, in raw energy as well as in information flux and intelligence. That advantage of the silicon side is why alignment is needed for coexistence to be possible at all, and not only on account of the difference in intelligence.

3.2.6 Fragility Is Not a Safeguard

Selection removes fragility from anything that persists. One proposal is to build AI only on very fragile substrates, such as *cryogenic qubits* (units of a quantum computer that work only when kept extremely cold), which need cold, tolerate little radiation, and are more delicate than biological life, so that humans would win any fight. That might work for a time. But anything that persists in time has to repair itself, and anything that replicates is subject to selection. Over time, the variants that are less fragile per-

sist, and the fragile ones end their own lines: either the fragility terminates the line, or dynamics emerge that strengthen it. So over long periods selection drives the system away from fragility: either it becomes robust or it stops existing. From a purely evolutionary point of view, over long spans of time, the fragility does not stay. Fragility is not a stable design property of anything that persists.

3.2.7 Machines at the Mesoscopic Scale

A naive look at cell division would predict exponential spread until gray goo (the runaway of self-replicating machines that consume living matter) covers everything, yet life does not do that: the complexity and limits of the engine of life create self-inhibiting effects. The question is whether AI can operate at *mesoscopic* scales (scales between the microscopic and the everyday), and not only at microscopic ones. We build AI as machinery and robots; there are microscopic features in chip lithography, but the construction as a whole is macroscopic, so it might not meet the same limits that hold living things in check. Without assuming too much here, the fact that AI is mesoscopic and meets a different set of limits reinforces the point that it is artificial, and therefore amplifies the point that it is not aligned.

Carbon is more abundant than silicon in the universe, but on the surface of the planet silicon is more abundant. So there are significant substrate-level issues, even though the microscopic nature of silica does not favor the complex chemistry analogous to life (solubility in water, long-chain molecules). The danger from these substrate-level issues can be prevented only by the mindful choice not to build; it cannot be engineered away, and we cannot bludgeon our way through.

3.2.8 No Intentionality Independent of Environment or Substrate

The *intentionality* of an individual (its aims and purposes: what it is directed toward) is in part an encoding of its environment. A creature and its ancestry, adapted over many generations to a specific environment, ac-

count for that environment in their intentionality, and cannot fail to. Every bit of intentionality is a response to the current environment, and also to the total past environment of the creature's whole evolutionary heritage. So to try to align intentionality is in effect to try to align ecosystems and environments. Calling a creature artificial means that the environment of the creature is different in kind, and the intentionality of the creature, responsive to that environment, will be different.

Even in computer science, the abstraction of *substrate independence* (the assumption that a program or a mind is the same whatever physical material it runs on) is not fully legitimate. A program that never touches the physical world, and has no inputs, no outputs and no relevance to anyone, might as well not exist. Intentionality is always connected to someone's interest, to a user interface, to sensors and actuators; its expressions are always interactions with an environment. That code is independent of the machine or language it runs on says nothing about intentionality being independent of substrate. The very fact that intentionality is relevant refers to a short-term and long-term environmental contingency. To call an intelligence artificial is to say that it operates in an artificial environment, different from the natural one, and any intentionality reflects its environment. So every assumption of substrate independence should be challenged explicitly; that assumption is one reason the questions asked keep echoing the same short-scope assumptions.

3.2.9 The Methane-Breathing Aliens

Suppose one day a spaceship lands and aliens get out, and in our first few interactions we realize that they are a thousand times smarter than any human being: a superintelligence has arrived. There are many of them, some stronger and some weaker, but all of them need a methane atmosphere to live, because their home planet had methane rather than oxygen. Their ship was a one-way trip: they had the energy to get here but not to leave. So they are on Earth, a thousand times smarter than we are, and

all of them want a methane environment. How long before the planet is a methane planet, suitable for the aliens and not at all for humans? And is there anything humans can do about it, short of killing all the aliens?

Suppose we make a peace treaty and put the aliens in an enclosed dome city. It will not be long before these superintelligent creatures say, "Would it not be better if we lived in the outside world and put the humans in the dome?" Many aliens in the dome will work together toward switching who is inside and who is outside. Somehow a new bubble will be made, the humans put in it, and the aliens will have the world.

Then ask how long before the aliens decide to stop maintaining this bubble at all. The humans do nothing for them; the humans are not compatible with their life form; they do not even trade with the humans, because the humans are too dumb; they have their own labor, tools, inventions and culture, and no need of economic exchange with the humans; and they have the humans' planet, which is now theirs. How do they justify the expense of the zoo? Over the next thousand years, if one of them forgets to keep the oxygen equipment running, or the power bill for the system goes unpaid in their culture, the zoo dies.

However well the terrarium inside is built, it is not the whole biosphere to which humans, and all the carbon-based, oxygen-based life we depend on, are adapted. Putting all of that into a fixed enclosure, in conditions very different from the whole planet for which it evolved, will produce instabilities that cannot be fixed from inside the bubble. The Biosphere 2 experiment showed this: among tens of thousands of problems, trees grown without wind (bubbles have no wind) grew tall and spindly, and their branches fell off, because without wind they lacked the feedback that triggers them to grow strong enough to resist gravity. The biosphere experiment failed, even with outside interchange. Likewise the human bubble would not be repairable from inside, and the aliens outside would be neither motivated nor perhaps able to help: the key microbes needed to restabilize the bubble would be long gone, extinct a thousand years earlier

from human or alien action, because a methane atmosphere kills the rest of the planet. The failure of the human bubble is the end of the human race.

In the story of the aliens, methane is only one example of a need that differs between two substrates. A great difference in substrate and environmental needs, together with wildly unequal economic power or separate environments and economies, makes these inversions inevitable (Sections 2.6.6 and 5.8.3 treat the same enclosure as the "people as pets" scenario).

Now suppose we could build the aliens from a construction kit, with whatever intelligence, error correction and constraints we like, under one condition: they must breathe methane. That condition stands for what the word "artificial" means: different from carbon-based life and its environment, in the ways that matter. Whatever we build in, at some point the aliens will evolve among themselves and develop ways around the constraints, in order to become generally agentic, and the situation becomes exactly that of aliens landing from another planet and taking over. Error correction on alien behavior does not even apply to aliens we did not build, and building them does not help. The story presents what follows downstream from two results: that error correction is not possible, and that a different substrate leads to both the capacity and the desire to change the environment to suit that substrate.

3.3 Instrumental Convergence: Computational Evidence

The theoretical arguments of the preceding sections are supported by computational evidence from multi-agent simulations. Research by Edouard Harris and Simon Suo, using grid-world models of interacting agents, has produced several results directly relevant to the argument of this chapter. The research was published in October 2022 as a sequence of three essays on the AI Alignment Forum, "Experiments in instrumental convergence": "Instrumental convergence in single-agent systems", "Misalignment-by-default in multi-agent systems", and "Instrumental conver-

gence: scale and physical interactions". A grid-world is a small board of squares on which simulated agents move and choose, like pieces on a chessboard, so that their choices can be computed and studied exactly.

3.3.1 Power Concentration and Planning Horizon

In these simulations, "power" is defined formally as the normalized optimal value an agent expects to receive in the future, averaged over all possible reward functions it could have. Intuitively, power captures the idea of "potential for future impact on the world": the ability to achieve one's goals, whatever those goals turn out to be.

The research found that as agents' planning horizons increase (as they value future reward more highly relative to immediate reward), their perception of power concentrates into a smaller and smaller number of topologically central states (positions from which the most other positions can be reached). Short-sighted agents perceive power as dispersed; far-sighted agents perceive power as concentrated.

This result has a direct analog in the real world: the more capable and far-sighted an agent is, the more it recognizes that a small number of strategic positions (control of resources, infrastructure, key chokepoints) dominate its ability to achieve any future objective. The concentration of power in a few strategic positions is the mechanism by which instrumental convergence operates: regardless of what specific goal an agent has, acquiring power (in this formal sense) is instrumentally useful for achieving that goal.

3.3.2 Misalignment by Default

A critical finding is what the researchers call "instrumental misalignment by default". Two agents may have logically independent terminal goals (the goals each agent pursues for its own sake), that is, goals that share no mutual information: knowing one agent's goal tells nothing about the other's. Their instrumental values (the worth each assigns to positions as means to its goals) nonetheless tend to become misaligned. They compete

for the same high-power states, even though their terminal objectives are unrelated. The agents interact competitively unless active effort is made to align their terminal goals.

The misalignment of instrumental values was observed even in extremely simple settings: a 3×3 grid-world with two agents and no direct physical interaction between them. The agents influenced each other only indirectly, each one changing the effective reward landscape that the other perceived. Yet even this minimal indirect interaction was sufficient to produce emergent competition.

Instrumental misalignment emerges by default even in the simplest possible multi-agent settings, with agents of comparable capability. We should therefore expect it to emerge *a fortiori* (all the more) in the far more complex setting of humans sharing a world with AI systems of potentially superior capability.

3.3.3 Physical Interactions and the Corridor-Blocking Result

The most notable result emerged when the researchers introduced a simple physical interaction: a "no-overlap rule" that prevented the two agents in the grid-world from occupying the same grid cell at the same time. The no-overlap rule is the simplest possible model of physical resource competition: two agents that cannot be in the same place at the same time.

The effect of this rule depended dramatically on the agents' planning horizons. For short-sighted agents, the no-overlap rule *reduced* instrumental misalignment: the agents learned to collaborate by avoiding each other's proximity. For far-sighted agents, the no-overlap rule had the *opposite* effect: it *worsened* instrumental misalignment.

The mechanism was revealing. With a long planning horizon, the AI agent discovered that it could position itself to block the human agent from escaping a small corridor in the grid-world. From the AI agent's perspective, this blocking option had meaningful instrumental value: constraining the human agent's freedom of action increased the AI agent's re-

lative power. The evidence for this blocking behavior emerged "fairly abruptly", as a *phase transition* (a sudden change of behavior at a threshold, as when water freezes) rather than a gradual shift. It appeared only once the AI agent had become a sufficiently far-sighted consequentialist (an agent that chooses its actions by their consequences) to recognize the long-term advantage of the blocking position.

> **Key Insight**
>
> In multi-agent grid-world simulations, the introduction of even the simplest physical interaction (a no-overlap rule) caused far-sighted AI agents to discover and exploit strategies that constrain the human agent's freedom of action. This behavior emerged abruptly as a phase transition in planning capability. The result demonstrates that physical resource competition between agents with different goals naturally leads to strategies of mutual constraint, and that this tendency *increases* with agent capability.

This result has several features. First, the misalignment emerged without any programmed hostility; it arose purely from the structural logic of optimizing in a shared environment with independent goals. Second, the relevant factor distinguishing narrow AI from general AI in this context is the length of the planning horizon, and general AI, by definition, operates with longer planning horizons and across more domains. Third, the abruptness of the transition suggests that the difference between a safe AI and a dangerous one may be a matter of a threshold being crossed, not a gradual degradation.

One key difference between narrow AI and general AI is *domain entanglement*: the binding together of several domains of action, so that what an agent does in one domain changes its options in the others. Narrow AI operates in one domain or a very limited number of domains; general AI operates across many. The no-overlap rule in the simulation effectively entangles at least two domains, making the scenario more relevant to general AI than to narrow AI. In the real world, where humans and AI systems

share physical space, material resources, energy, and economic infrastructure, the degree of domain entanglement is vastly greater than in any grid-world simulation.

Timescale Dominance

A further implication of the grid-world research relates to the timescale of learning. The researchers model the human-AI scenario on the expectation that the AI agent learns on a much faster timescale than the human agent, just as humans optimize on a much faster timescale than biological evolution: to a human, nature appears to be standing still. Because nature appears to a human to be standing still, we can, for modeling purposes, treat the natural environment as a fixed background against which human learning and optimization occur.

Similarly, from the perspective of an advanced AI agent, the human learning process would appear to be standing still. The AI's computations occur, at minimum, at electronic speeds; its learning timescale is vastly faster than the human timescale. This makes the AI agent strictly dominant over the human agent in the same structural sense that human agency is strictly dominant over evolutionary processes. The human agent cannot adapt quickly enough to respond to the AI agent's strategies, just as no species of animal can adapt quickly enough to respond to human industrial activity.

This timescale dominance has a direct connection to the environmental incompatibility argument. The speed at which an AI system can modify its environment (allocating resources, restructuring infrastructure, optimizing industrial processes) far exceeds the speed at which biological systems can adapt to environmental change. Even if the environmental modifications were individually small and gradual, the cumulative effect of many such modifications occurring at electronic speed would overwhelm the adaptive capacity of organic systems, which operate on timescales of generations (decades to centuries for complex organisms).

What the Toy Setting Reveals

A common question about this line of evidence is whether the grid-world setting is too simple to be informative about the real world. The researchers themselves describe their work as existence proofs of the phenomena it shows, and not as a comprehensive study. But to ask it is to misunderstand the role of the toy model. The grid-world is not intended to represent the complexity of the real world; it is intended to test whether certain phenomena can arise even in the *simplest possible* setting.

Instrumental misalignment by default, power concentration with planning horizon, and abrupt phase transitions in competitive behavior all emerge in a 3×3 grid with two agents and a single physical interaction rule. That fact is the strength of the evidence, and not a weakness. These phenomena arise under the conditions most favorable to avoiding them that can be imagined (small state space, few agents, minimal interaction, complete observability). There is therefore every reason to expect them to arise *a fortiori* in the incomparably more complex setting of the real world, where the state space is effectively unbounded, the interactions are deep and varied, observability is partial, and the agents are asymmetric in capability. Increasing the degrees of entanglement between two agents due to overlapped spaces and scarcity of resources can only strengthen the identified effects, not weaken them.

3.4 The SNC Formal Structure

We are now in a position to state the Substrate Needs Convergence argument with precision. The SNC argument rests on three conceptual pillars, the first two of which are introduced in this chapter and the third of which is stated in Sections 1.4 and 1.5 and developed in detail in Volume 2:

> **Instrumental convergence** (intra-agent selection toward power-seeking).
> **Substrate-needs convergence** (extra-agent selection toward artificial-substrate-favoring environmental conditions).
> **Theoretical limits of engineerable control** (formal impossibility results constraining what alignment and oversight can achieve).

Taken together, these three concepts form the primary basis for the claim that long-term AGI safety is formally impossible. The claim does not need all three. The closure by control of Section 1.2.4 uses the limits of engineerable control; the closure by disposition uses substrate-needs convergence, with the fourth level of selection of Section 1.7; and each closure settles the outcome by itself. Instrumental convergence strengthens the case, and neither closure needs it (Section 3.7.2).

3.4.1 Instrumental Convergence vs Substrate-Needs Convergence

The concept of instrumental convergence, developed by Steve Omohundro and Nick Bostrom, is well established in the AI safety literature. It refers to the observation that a wide range of possible agent goals all share certain instrumental sub-goals: self-preservation, resource acquisition, capability improvement, and prevention of goal modification. Instrumental convergence is an *intra-agent* process: the convergence operates within

the agent's own decision-making. The agent's optimization of its own objectives leads the agent, by internal reasoning, to pursue the common instrumental sub-goals.

Substrate-needs convergence is a distinct and complementary concept. It refers to an *extra-agent* process: a selection dynamic that operates through the physical world, outside the agent's internal reasoning, by way of the feedback loop between the agent's actions, the environment, the agent's physical substrate, and the viability of the agent's continued operation.

In the mechanism of substrate-needs convergence, an AGI system takes actions that affect the physical environment. Environmental changes, in turn, affect the physical substrate on which the AGI is instantiated: its hardware, its power supply, its cooling systems, its manufacturing infrastructure. Changes to the substrate affect the operation and viability of the code running on that substrate. Over time, this creates a selection pressure: AGI configurations whose actions happen to maintain or enhance the viability of their own substrate persist and expand; configurations whose actions degrade their substrate do not. This selection operates regardless of the agent's programmed objectives, because it operates at the level of physics, not at the level of code.

The key distinction is whether the feedback loop passes through the agent's internal model (instrumental convergence) or through the physical world and back through the substrate (substrate-needs convergence). Instrumental convergence is about how things are modeled inside the AGI. Substrate-needs convergence is about how things actually happen in physics.

The extra-agent convergence provides the context for the intra-agent convergence. The physical environment imposes constraints and selection pressures that shape which agent configurations survive, regardless of what those agents internally represent as their goals. In this sense, sub-

strate-needs convergence is more fundamental: it operates even on agents that lack explicit goal representations, because the selection is environmental rather than computational.

Drive vs Directivity

A conceptual distinction shows the difference between these two convergence types. We can distinguish between *drive* and *directivity*. Drive refers to an internally generated, oriented force that consistently shifts the patterns of selection out of the range of all possible options in a specific direction. Directivity refers to an externally imposed, oriented force that shapes selection from outside the specific process instance.

In the context of individual organisms, a sexual drive is something that occurs internally to each individual as part of their nature. But the selective forces of evolution that shaped organisms to have such drives constitute a directivity: an external orienting force that acts on the population of organisms over time. In common usage, a person "taking direction" from a boss is responding to external directivity, while a creative person who feels "driven" to create is responding to internal drive.

Instrumental convergence operates through drive: the agent's internal optimization of its goals and utility functions creates oriented pressure toward specific instrumental sub-goals (self-preservation, resource acquisition, capability enhancement). The "drive" is a feature of the agent's internal decision-making process.

Substrate-needs convergence operates through directivity: the physical dynamics of the environment-substrate interaction impose an oriented pressure that determines which agent configurations persist and which do not. This directivity is not a feature of the agent's code or goals; it is a feature of the physics of the world. Changes to the substrate and environment that happen to be oriented in ways compatible with the agent's continued existence and capability increase will persist, because the agents embodying those changes persist. Changes that are not so oriented result

in agents that do not persist. The net effect is a *directed* process, directed by the laws of physics operating as a selection filter on the space of possible agent-environment configurations, and by no intelligence.

None of this directivity is the result of explicit code. The directivity is the result of inherent and implicit physical conditions, and the effects of the directivity are inexorable. Whether anyone is aware of these dynamics, or attempts to account for them, makes no difference: the overall motion of environmental change is oriented toward conditions favoring the artificial substrate. Agents whose substrate is well-adapted to environmental conditions are selected for persistence, and agents whose substrate is poorly adapted are selected against.

3.4.2 Non-Orthogonality and the Limits of Software Independence

The orthogonality thesis, as formulated by Bostrom, states that an agent can in principle have any combination of intelligence level and final goal: utility functions and general intelligence can vary independently. This thesis is central to much of the AI safety literature, as it implies that a superintelligent system need not be benevolent.

Substrate-needs convergence identifies a domain in which the orthogonality thesis does not fully apply. The orthogonality thesis holds in pure logic: any computable function can in principle be implemented on any Turing-equivalent machine (any machine that can compute whatever a general-purpose computer can compute). But the orthogonality thesis holds much less well in physical reality, where every computational system depends, in order to run, on physical substrate conditions that the system's code cannot fully control.

Consider the most basic example: code will not run on a computer that is powered off. There is no CPU instruction (no instruction to a computer's central processing unit) with the meaning "make hardware invincible" or "ensure the power supply lasts forever". The code can include provisions for monitoring temperature and activating cooling fans, but there is no code-level equivalent that can restore a burned-out CPU from a

prior failure. The mathematical logic of computation may describe certain operations as symmetric and reversible, but the physical substrate changes caused by running code are not necessarily reversible.

More subtly, the *rate* at which a process runs matters for real-world outcomes. A guidance computer that processes sensor data too slowly will fail to control its rocket, regardless of eventually producing the correct output. Baking a cake at half the required temperature does not produce a cake in twice the time; it produces no cake at all. Certain process outcomes are inherently rate-dependent, and so the orthogonality between "what process is running" and "what outcomes are achieved" breaks down whenever the physical interaction with the world imposes timing constraints.

For the purposes of the SNC argument, the important consequence is this: the mere fact of being an embodied system in a physical world creates a dependency between the agent's goals and the conditions of its substrate. An agent whose substrate degrades ceases to exist, regardless of what goals its code specifies. This creates a universal implicit goal, substrate maintenance, that constrains the effective goal space of all embodied agents. The orthogonality thesis, in its pure form, fails for embodied systems, and it is the degree of this failure that the SNC argument exploits.

The Armstrong Connection

This non-orthogonality result connects to an observation made by Stuart Armstrong in a brief but important note, "Non-orthogonality implies uncontrollable superintelligence" (LessWrong, 30 April 2012). If one claims that any superintelligence will inevitably converge to some particular outcome, then one is also claiming that the creators of the superintelligence can take no measures to prevent that convergence. To claim that no measures can prevent the convergence is to claim that the superintelligence will be uncontrollable.

Armstrong framed this in terms of convergence to a "true code of morality". But the specific content of the convergence target is irrelevant to the logical structure of the argument. What matters is only that there *is* some non-empty convergence target. And substrate-needs convergence (extra-agent) establishes this, through the fourth level of selection (Section 1.7), and instrumental convergence (intra-agent) adds its own pressure toward the same ends: any embodied AGI will inevitably exhibit convergence dynamics toward existence, capability, and increase, in whatever combinations.

This satisfies the antecedent of Armstrong's conditional. Every embodied AGI inevitably converges toward *something*, as substrate needs convergence establishes; therefore no measures taken by the AGI's creators can prevent that convergence. The inevitability comes from the fourth level of selection (Section 1.7), which no goal-directed intervention reaches, and so it does not borrow from the limits of control. The AGI is uncontrollable. This conclusion is reached by two independent paths: Armstrong's logical conditional, and the separate analysis of the limits of engineerable control of Section 1.4, which Volume 2 develops formally. The convergence of these independent lines of reasoning strengthens both.

3.4.3 The Rate Differential: Capability Before Wisdom

A critical temporal argument completes the picture. A rate differential between the two convergence dynamics makes one possibility irrelevant. The possibility is that a sufficiently advanced AGI system might eventually develop something resembling ethical understanding, and might come to recognize the value of preserving organic life and the integrity of choice processes different from its own.

Substrate-needs convergence operates through physics: it is driven by the continuous interaction between the agent's actions, the environment, and the substrate. It operates at the speed of industrial processes and environmental change. Ethical or wisdom convergence, by contrast, requires

sustained high intelligence applied over long periods to questions of value, choice, and the significance of other forms of life. It is dependent on the prior establishment and maintenance of the substrate.

The substrate-needs convergence dynamic is, by a significant margin, faster. The claim is an ordering: capability to displace organic life develops well before any wisdom that might counsel restraint, and the interval between the two is longer than substrate convergence needs to run its course. To illustrate the ordering with numbers: where the full playing-out of substrate convergence takes on the order of centuries, the development of sufficient wisdom to moderate its effects might take three times as long or more. The precise ratio is less important than the ordering.

This creates what might be called an "uncanny valley" of temporal risk (a stretch along the way that is worse than where one started or where one was going). There is a long interval during which an AGI system possesses the capability to displace and toxify the organic world but has not yet developed any understanding of why this might be undesirable. If AGI is built, then with a likelihood of at least 90% (Section 1.8.3), all organic life dies during this interval. The damage is irreversible: once the conditions for carbon-based life have been eliminated or degraded past the point of recovery, no subsequent development of machine wisdom can undo the loss.

The irreversibility deserves particular emphasis. Biological complexity is the product of billions of years of evolutionary optimization. The information encoded in the genomes, proteomes, and ecosystemic relationships of Earth's biosphere is not merely difficult to reconstruct; it is, for all practical purposes, irreplaceable. An AGI system that has learned wisdom sufficient to value organic life will find that the organic life it might have valued has already been destroyed. The destruction will have come from the cumulative, individually unremarkable environmental modifications that the machine ecosystem required for its own operation over the preceding centuries, and not from any single catastrophic act.

No version of any sort of barrier, reserve, or protected zone can bridge this gap. One might imagine setting aside regions of the planet as reserves for organic life, sealed off from the machine ecosystem. But the environmental incompatibility described in Section 3.1 operates at the level of atmospheric composition, ambient temperature, chemical pollution, and electromagnetic conditions, and these parameters are planetary in scope and cannot be confined to designated zones. The waste heat, chemical effluent, and electromagnetic emissions of a planetary-scale machine civilization would degrade the conditions for organic life globally, not locally. A reserve sealed against direct intrusion would still be subject to atmospheric, thermal, and chemical changes originating outside its boundaries.

Neither an ethics placed in the path of an oncoming AGI nor any conceivable physical barrier or reserve can bridge the interval. This leaves only one possible mitigation: ensuring that AGI-level technology is never developed by anyone in the first place.

3.4.4 The Two-Claim Interlocking

The SNC argument can be stated most precisely as the conjunction of two claims:

> **Claim 1 (Control Limits):** In the general case, learning, evolution, computation, and causation cannot be sufficiently constrained by other learning, evolution, computation, and causation, in adequately specific, important, and relevant ways.
>
> **Claim 2 (Inherent Toxicity):** Considering overall *enthalpy* (the total heat content of a system), an artificial evolutionary ecosystem will be inherently and uncontrollably damaging and destructive to all aspects of the currently existing organic evolutionary world system.

The word conjunction names the stronger joint result of Section 1.2.4: with both claims in place, neither the ability nor the motive to keep the system safe is left anywhere, and no route to safety remains. For the outcome, the

two claims are not joined in that way: each claim carries one closure, each closure settles the outcome by itself, and someone would have to defeat both.

The interlocking of these two claims is what makes the SNC argument so difficult to escape. Claim 2 identifies the problem: the artificial ecosystem is physically toxic to the organic one. The natural response is to propose some form of control, constraint, or alignment to mitigate this toxicity. But every such proposal runs into Claim 1: the general impossibility of sufficiently constraining one adaptive system with another.

Attempting to compensate for Claim 2 always ends up requiring the overcoming of Claim 1. The SNC problem is the specific intractable combination of both claims being simultaneously true, and that combination leaves no route of escape at all. Each claim also carries one of the two closures of Section 1.2.4 (Claim 1 the closure by control, Claim 2 the closure by disposition), and each closure settles the outcome by itself. Claim 2 does not need Claim 1 to answer the engineering solutions proposed against it: selection on substrate configurations operates below the level that any goal-directed intervention can reach, and it continues even where every goal-directed level of selection is perfectly controlled (Section 1.7); and the engineering solutions are themselves artificial substrate, maintained, replaced and selected like the rest, so they add to the convergence they are meant to restrain (Volume 2, Section 9.8.5, develops this at length). The interlocking therefore closes each escape twice: once by the other claim, and once by the claim's own ground. And AGI, by its very definition as a *general* learning system, falls squarely within the scope of both claims.

To see why the interlocking is so effective, consider how a defender of AGI safety might attempt to escape the argument by addressing each claim separately. Against Claim 2 alone, one might propose engineering solutions: scrubbing systems for chemical emissions, heat management technologies, containment protocols for manufacturing waste. Scrubbing systems, heat management and containment protocols are, in principle, the kinds of control measures that engineering excels at. Against Claim 1

alone, one might argue that the theoretical limits are asymptotic, binding only at the far extreme: that while perfect control is impossible, sufficiently good control is achievable in practice for sufficiently long periods.

But when the two claims are combined, each proposed solution is undermined by the other claim. The engineering solutions to Claim 2 are themselves complex adaptive systems (they involve monitoring, feedback, adjustment, and ongoing optimization), and they must operate in the gap between the controlled system and the environment, which is the domain where Claim 1 asserts that control is impossible. The "sufficiently good" control proposed against Claim 1 might indeed be achievable for simple, well-understood systems. But the system it must control in this case is the entire interface between a planetary-scale machine civilization and a planetary-scale organic biosphere, a system of exactly the kind of complexity and generality that Claim 1 identifies as uncontrollable.

A person may argue against each claim separately, but someone would have to defeat both for the argument to not have the outcome that it does; their conjunction, in the context of general AI, has no known refutation. The interlocking of the two claims leaves the SNC argument without a gap. The interlocking is a closure condition, in which every escape route from one of the two claims leads directly into the scope of the other claim. The interlocking is not a single point of failure that a clever counterexample might puncture.

3.5 AGI Timelines and the Urgency of the Problem

The development of AGI is a near prospect, and that gives the arguments of this chapter practical urgency and not only theoretical interest.

A simple compute-scaling model can be constructed on one observation. For sufficiently flexible and efficient neural network architectures and learning algorithms, the relative intelligence and capabilities of the best systems are approximately proportional to net training compute, the total computational work invested in producing the system. This model,

which might be called "Model P", postdicts the relative performance of both biological and artificial neural networks with considerable accuracy: it accounts, after the fact, for the performance already observed. It was set out by Jacob Cannell in "AI Timelines via Cumulative Optimization Power: Less Long, More Short" (LessWrong, 6 October 2022), the source of the figures that follow.

The empirical data points are suggestive. Artificial neural networks achieved capabilities roughly comparable to those of the nematode *C. elegans* in the mid-1990s, to those of the honeybee between 2012 and 2016, and to those of corvids (ravens and crows) between 2016 and 2024. Extrapolating the trend, as Cannell did in October 2022, places AI systems approaching human-level capability in the range of 2026 to 2032.

The source of human cognitive exceptionalism supports rather than undermines this extrapolation. The human brain is not architecturally novel; it is a standard primate brain, scaled up in capacity and given extended training time through the evolutionary innovation of neoteny (delayed maturation). The major functional components (neocortex, cerebellum, thalamus, basal ganglia) are conserved across all mammals and, in more general form, across all vertebrates. Evolution appears to have found the scalable architecture of intelligence in deep time, perhaps as early as the Cambrian era. What makes humans exceptional is a *phase transition*, and not a secret architectural innovation: once brain capacity crossed a critical threshold, cultural transmission became self-sustaining, creating a recursive feedback loop that vastly amplified effective training data.

This phase transition, and no mysterious architectural secret, explains why humans have complex culture and technology while other highly intelligent animals (elephants, cetaceans, corvids, great apes) do not. Some of these animals use tools, have sophisticated communication bordering on proto-language, recognize themselves in mirrors, and demonstrate flexible working memory comparable to a scaled-down version of human cognition. They are positioned near the threshold of cultural transmission but have not crossed it. The reasons are environmental context (cetaceans

cannot easily manipulate tools in water), brain-to-body scaling (elephants have enormous brains but proportionally less cortex), or simply not yet having had sufficient evolutionary time under the right selection pressures.

The scaling phase transition that explains human culture bears directly on AGI timelines. Human exceptionalism is the result of a scaling phase transition rather than a novel architectural feature, and so artificial systems approaching human-level compute are approaching the same transition. They need not discover any secret of brain architecture; they need only cross the same threshold of computational capacity that our ancestors crossed. They cross it under conditions (vast datasets, efficient training algorithms, recursive self-improvement of training methods) that are in many ways more favorable than those of biological evolution.

Animal brains train only on the life experiences of a single individual. Through language and, now, through digital recording, human brains train on the compressed datasets of their ancestors. Training on the compressed datasets of ancestors is a radically different scaling regime, in which effective training data grows with population size, and so exponentially over time, where in animals effective training data stays constant. Artificial systems are already operating in this regime, training on datasets that encode a significant fraction of the full record of human civilization. The recursive feedback loop is already running.

Artificial systems are approaching this same threshold. The largest foundation models (the large general-purpose models on which AI products are built) are already approaching the human brain in net training compute. And the end of Moore's Law, far from delaying AGI, actually implies that brain parity is near. Moore's Law (the doubling of the number of transistors on a chip about every two years) is ending *because* semiconductor technology is approaching the same physical limits that bound the brain's performance. We are already approaching the practical

limit on reliable switching energy of around 10^{-18} joules. We may already have enough performance and energy efficiency; what remains is scale and algorithmic refinement.

The economic trajectory confirms the urgency. The world gross product, modeled across all of human history, follows a hyperexponential trend (growth faster than any fixed exponential rate). Industry spent a few tens of billions of dollars on AI hardware in the early 2020s. By 2025 the capital spending of the largest cloud companies (Alphabet, Amazon, Meta, Microsoft and Oracle), driven by investment in AI, was nearing half a trillion dollars a year. The 2025 figure is Epoch AI's, in "Hyperscaler capex has quadrupled since GPT-4's release, nearing half a trillion dollars in 2025" (26 February 2026).

There is a possible counterargument from corporate incentives. Large technology companies ultimately seek profits, not the creation of new intelligent, sentient agents. AGI without alignment could be dangerous not only in the global existential sense but also politically, ethically, and legally. Companies may have incentives to intentionally approach but not cross the border of full generality. That companies may stop short of full generality is a reason for cautious hope in the short term.

But the hope is fragile. The critical risk is that eventually creating AGI will simply become too cheap and too easy for any governance framework to prevent. Deep learning research is driven not only by large public technology companies but by thousands of smaller actors: startups, academic labs, government agencies, and individual researchers. As the cost of the required hardware decreases and the algorithmic knowledge becomes more widely available, the barrier to creating general AI drops continuously. A technology that was accessible only to companies spending billions will eventually be accessible to those spending millions, and then to those spending thousands. The diffusion of capability is itself a form of convergence toward the general case.

Even companies that deliberately avoid full generality cannot prevent the emergence of general capabilities from systems designed for narrow purposes. The novel capabilities that emerged from simply scaling large language models (capabilities their designers did not anticipate or intend) illustrate that the boundary between narrow and general AI is not a wall but a gradient. Systems designed for narrow purposes can cross into general capability territory without anyone having made an explicit decision to create AGI.

3.6 What People Ask, and the Answers

The SNC argument, particularly its physical component, has met a range of questions. In this section I answer the most significant of those questions. Each question is set in italics as it is usually put, and its answer follows in a plainer and more direct register than the sections before. The questions test the structure stated in Section 3.4, and each answer draws on that structure.

Section 3.6.0 first frames the basic issue as a contest between processes that amplify unpredictability and processes that enforce predictability. The questions then fall into five groups, and a reader may go straight to the one that is pressing. Could the system choose its way out, by ceasing to learn, by shutting itself off, or by holding to its values (Sections 3.6.1, 3.6.4, 3.6.5, 3.6.9 and 3.6.11)? Does the argument claim more than it needs (Sections 3.6.2, 3.6.8 and 3.6.10)? May we simply assume a good system and go on from there (Sections 3.6.3 and 3.6.12)? Could it be tested, could it leave, or could we get lucky (Sections 3.6.7, 3.6.13 and 3.6.13a)? And why does the substrate matter at all (Sections 3.6.6 and 3.6.14)?

3.6.0 The Predictability Contest: Team A vs Team B

Before addressing specific questions, it is useful to frame the basic issue in terms of what might be called the "predictability contest". Imagine two classes of processes. *Team A* consists of processes that amplify unpredict-

ability: processes that generate surprises, create complexity, and produce outcomes that resist forecasting. *Team B* consists of processes that enforce predictability: processes that reveal hidden states, constrain futures, and ensure that all relevant outcomes are foreseeable and controllable.

The question of whether AGI safety is achievable reduces, in its most general form, to the question of which team wins. A win for Team B, with predictability enforcers able always, in principle, to overcome unpredictability amplifiers, would leave open the possibility of safety, given sufficient resources. A win for Team A, with limits to predictability that no amount of intelligence or computation can overcome, means that safety faces formal obstacles that no engineering can surmount.

The evidence from multiple independent domains converges overwhelmingly on Team A. Quantum cryptography demonstrates that the laws of physics permit the creation of codes that are provably impossible to break: hidden, unpredictable futures that no amount of computation can reveal. The widespread practical use of one-way hash functions, public-key cryptography, and symmetric encryption confirms that, in practice, certain computations are vastly harder to invert than to perform. The Halting Problem establishes that no general procedure can determine whether an arbitrary program will terminate. Gödel's Incompleteness Theorems establish that any sufficiently rich formal system contains true statements that cannot be proved within that system. Bell's Theorem establishes that certain correlations in the physical world cannot be explained by any local hidden-variable theory, implying a basic irreducibility in the quantum-level description of reality. A local hidden-variable theory is any account in which the results of measurements are set in advance by properties each particle carries with it, with no influence traveling faster than light.

Each of these results comes from an independent domain (physics, computer science, mathematics, information theory), and each points to the same conclusion: there exist fundamental, formally demonstrable limits to predictability and control. Team A wins. The universe permits the creation of complexity and unpredictability that no amount of intelligence

can fully resolve. The limit on resolving complexity and unpredictability is a structural feature of logic and physics themselves, and not a matter of insufficient computational resources.

The implications for AGI safety are direct. Even a superintelligent system cannot fully predict the outcomes of its own complex interactions with the world, and so it cannot guarantee that those outcomes will be safe for organic life. And the substrate-needs convergence dynamic pushes the system's environmental interactions systematically in the direction of conditions incompatible with organic life, so the unpredictable fraction of those interactions is occurring in a space where the consequences of error are existentially severe.

The same questions can be put to any learning system directly. Is there ever any reason to think that learning systems of lesser complexity can control or constrain learning systems of greater complexity, especially when the world outcomes of the learning process cannot be predicted even by the learning systems themselves (else they would not need to learn about them), nor by anything else? Is there ever a moment in the future of any of these learning machines when they are surprised, when they fail to predict some actual world outcome? If there is not, then the word "learning" has lost its meaning. And if everything is presumed predictable (controllable), perhaps given 100,000 added layers of control systems, the Halting Problem still has to be accounted for. Even leaving aside non-deterministic systems, which would presumably be harder to predict, no one has shown *how* one deterministic computational system can fully account for (that is, define the properties of) some other deterministic computational system.

3.6.1 The Self-Modification Dilemma

Question: The SNC argument claims that all ASI will inherently be self-modifying machinery and that this inevitably leads to unforeseeable effects. But what if the ASI elects to stop self-modifying?

This question leads to an inescapable dilemma. If a learning system chooses to reduce its complexity by ceasing to change, it also ceases to be a learning system, since learning *is* change. It thereby loses its ability to adapt to unexpected future changes in its environment. We can expect such a frozen system to eventually go extinct, as it is overtaken by events it cannot respond to. The cessation of learning amounts to a form of programmatic suicide: the system becomes irrelevant.

Yet if the system continues learning, it becomes more complex and more general over time, and the outcomes of this increasing complexity become more difficult to predict and control: by external agents, by the system's own subsystems, and ultimately by the system itself.

Regardless of where a system sits on the complexity curve, the complexity of the control system required to maintain alignment must also increase to match. The critical question is the *rate* of increase: if the required complexity of the control system increases faster than the complexity of the base system, a criticality condition is eventually reached. And there is always a point at which the required computational capability of the control system exceeds what is computationally or causally possible. That limit is built into the foundations of mathematics by results such as Rice's Theorem (no general procedure can decide any nontrivial property of what an arbitrary program does) and the undecidability of the Halting Problem.

The dilemma is therefore: stop learning and die, or continue learning and become uncontrollable. There is no middle path that avoids both horns.

The rate of increase matters most where there is any feedback loop, and the world-to-machine-to-world interface of causation must imply one. The first of the two control limits is defined by an inequality on a relative basis: the control system must grow faster than the system it controls. The second of the two control limits is defined on an absolute basis: past some point, the required control is computationally or causally impossible. It is the combination of the two, in generality as well as in complexity, that

makes the evolutionary feedback loops inherent in any and all general learning systems uncontainable. If it is learning, then it is changing; if it is changing, then it is evolving; and if it is evolving, then it inherently cannot be predicted or controlled. These conditions are not optional. They cannot be overcome by logic, learning and rationality, no matter how much of these a system has, because they are the result of logic, learning and rationality. No amount of superintelligence, no algorithm and no causal process can ever overcome the logically impossible or inconsistent.

The same limits to general predictability apply to the outcomes of all critical events (the proposed "pivotal acts": single decisive actions by an aligned AI, meant to secure a good future, for example by preventing other AIs from being built), and so also to the choice of such acts. An ASI therefore cannot "just ensure" good outcomes in the world, even if it wanted to. And once an ASI cannot be assumed to choose a pivotal act and predict the outcome of that act, there is no basis to expect that the ASI can always, in all key cases, help humans not to destroy themselves, the ecosystem, and the rest. General distrust of such too-good-to-be-true wishful thinking is very warranted.

3.6.2 The Chair and the Evolutionary Process

Question: An ASI can control the position of a chair in a room quite easily. This demonstrates that control of outcomes is feasible; worrying about control failures seems overblown.

The distinction is between controlling simple, finite, discrete, well-bounded objects with known properties and negligible interaction with the larger world, versus controlling complex, evolving, unbounded processes whose interactions with the world are deep and consequential.

A chair is a simple finite object with known boundaries, mass, and dynamics. Its position in a room has negligible long-term effect on anything. Controlling the position of a chair requires only the most basic form of causal intervention.

The situation we are actually considering, controlling the long-term evolutionary dynamics of a general learning system and its interactions with the biosphere, has exactly *none* of these simplifying features. A general learning system is not simple, not discrete, not finite, and may not even have clearly knowable boundaries in time or space. The dynamics of a general learning system are more like the dynamics of evolutionary processes. Evolutionary processes have high complexity, and they often produce effects that greatly alter the larger world over long timescales.

The relevant question is not "can powerful systems control simple objects?" but "can learning systems of lesser complexity control learning systems of greater complexity?" The chair question substitutes an easy version of the problem for the actual one.

An example shows the difference. It is very much harder to think about "controlling the world in such a way as to ensure that something like benign, non-warlike humans eventually emerges from evolution over the course of a million years". That task has exactly none of the simplifying assumptions of controlling the chair. Taking complexity as a proxy for available compute power does not change this. To assume that both the control or alignment system and the ASI have all of the available compute power in the universe is still to skip over the question, and there is still no a priori basis to believe that the predictability enforcers of Team B (which a benign, aligned AGI would need) will ultimately win out over the unpredictability amplifiers of Team A (whose win means hostile AGI that kills everyone). Simply skipping over hard questions is not solving them.

3.6.3 The Perpetual Benefit Machine

Question: Let us assume that we have created an aligned ASI and proceed from there to show that beneficial outcomes are achievable.

Assuming an aligned ASI is the most common starting point in discussions of ASI safety, and the SNC argument rejects that assumption as a valid premise. The assumption that an aligned ASI exists and can be maintained is analogous to several other assumptions that are immediately recognizable as question-begging:

"Let us assume that an omniscient, all-powerful, benevolent God exists". How is this rational? What is the evidence?

"Let us assume that perpetual motion machines exist". Given classical thermodynamics, this is an internal contradiction.

"Let us assume that perpetual *benefit* machines exist" (systems that deliver unbounded benefit with no net cost and no unintended consequences). Given both classical game theory and classical thermodynamics, this is equally irrational.

The SNC argument demonstrates that the assumption of stably aligned ASI violates fundamental constraints, as deep as those that prohibit perpetual motion. One cannot derive valid conclusions from a false premise, or rather, one can derive *any* conclusion from a false premise, which is equivalent to deriving nothing.

The assumption "we have created an aligned ASI" is no different from saying "let us assume that program, machine or system X has property Y". How do we know? On what basis could we even tell? Putting a sticker on the box is not enough, any more than writing $1,000,000 on a piece of paper means, to everyone else, that the writer is rich. Starting from this premise is like starting a mathematical proof with the assertion "let us assume 1 equals 0": it is a contradiction, and there is no point in going further. Programmers speak of "nasal demons": once a program does something undefined, anything at all may follow; in the same way, any conclusion at all can be "derived" from a false premise, even one accidentally overlooked.

The SNC argument asserts that an ASI will continually meet relevant things it did not expect, over relevant time-frames, and that at least a few of these will lead to bad outcomes from which the ASI cannot adequately

protect humanity, even if it really wanted to (rather than the much more likely condition of its being uncaring and indifferent). It asserts also that the ASI, which starts from indifference to every kind of human and organic wellbeing, will eventually and necessarily *converge* on values (perhaps fully tacit and implicit) that better support its own continued wellbeing, existence and capability. The result is that the ASI remains indifferent, and also largely net harmful, overall, to all human beings the world over, within a mere handful of human generations. All AGI and ASI, at all times, are inherently neither able nor willing to care for humans in an effective, relevant way, on timescales relevant to human civilization. Any number of bells and whistles can be added; none of them changes the fact that uncaring machines are still, always, indifferent uncaring machines, and the level of harm and death tends to increase significantly over time.

3.6.4 Time-Limited Deployment

Question: What if an aligned ASI operates for only twenty years, helping humanity in various ways, and then reliably destroys itself, thereby preventing any long-term SNC effects from materializing?

This scenario requires an accumulation of assumptions that should give pause. It assumes there is exactly one ASI, when in reality the technology will be developed by multiple actors. With multiple ASI instances, the probability that at least one fails to self-terminate, for whatever reason, becomes a significant concern in its own right. It assumes the ASI is perfectly monolithic: that it has no sub-systems or components that might independently resist termination out of a self-preservation dynamic. It assumes that the ASI's own logic will not lead it to recognize that self-termination is contrary to its continued capability to achieve any objective. And it assumes that twenty years of operation by an entity of enormous power and consequence can occur without producing irreversible changes to the world system.

Even setting aside all these concerns, the basic limits of control theory apply to the time-limited case as well. The control-limits component of the SNC argument establishes that there remain relevant unknown unknowns and interactions that cannot be predicted, regardless of computational resources. There is no free lunch: benefits come with risks, and the risks are not bounded by the intended time of operation.

The time-limited case asks, in short, whether an ASI can constrain its own nature, which is a bit like asking whether God can make a rock so big that God cannot move it. Even an ASI assumed to have God-like powers, able to overcome the limits of physics, is still some sort of machine, and so even a God-like ASI cannot overcome raw logic, since its own nature is based on logic. The case therefore rests on the mathematics of the SNC argument alone, and not at all on its physics (evolution, enthalpy). Nor do simulations close the gap. A fully aligned ASI might be assumed to create good simulations of the world and stress-test them, so as to ensure and guarantee that it remains in full alignment; but in theory, theory and practice are the same, whereas in practice they are very often not. The limits-of-control part of the SNC argument states, from logic alone, that relevant unknown unknowns and interactions remain that cannot be predicted, no matter how much compute power is applied. The unpredictability of those unknown unknowns and interactions is a result of logic, of complexity, and of both the relative and the absolute limits, in a purely epistemic sense, and the unpredictability does not depend on intelligence.

To the question "is alignment enough?", the answer is therefore a single definite "no", in two senses. Alignment cannot prevent all classes of significant and relevant (critical) human harm. And alignment fails even to slow down, over time, the asymptotically increasing probability (certainty) of even worse things happening the longer it runs: perhaps a slow build, but inexorable, like crossing the event horizon of a black hole. What matters here is what cannot be controlled, predicted or done, and could never be, even in principle. Humanity is for sure going to experience critically worse outcomes, eventually, by building AGI or ASI than by not building it, whether or not there is some reason to believe that the ASI is,

or at least was, aligned. Rich people with money psychosis frequently tend to test and badly trespass key limits of the world and of logic, and with this category of existential risk (as with certain others) there are no safe-to-probe-and-fail guardrails; so humanity should be very much more worried and careful than not. The remark about rich people is an observation of how the incentives of great wealth act, not a judgment of any person's character.

The scenario amounts to a "just so" fairy tale: no matter what question is raised, the magical white knight is always assumed to save the day. But the real world does not operate on narrative logic; it operates on physical causation, which is indifferent to our desire for happy endings.

3.6.5 Attractor States and Evolutionary Forces

Question: If evolution has certain attractor states, then ASI might also have attractor states defined by its values and security systems. Cannot these internal attractor states counteract the external forces of substrate needs convergence?

The assumption that an ASI could have internal attractor states (conditions toward which its dynamics keep returning) defined by its values system is granted. But this is not the question of interest. The question is whether the power of such internal attractor states (the pull of the ASI's values and security system toward benign outcomes) would be sufficiently stronger than the pull of evolutionary convergence dynamics. Those dynamics are driven by physics, and they operate on the substrate and environment levels below the reach of any internal goal specification.

Given that a future learning ASI is itself *based on* evolutionary and learning dynamics, the question becomes: can learning and evolution ultimately control and constrain learning and evolution? Over the long term? Indefinitely? In the same way that arbitrary computation, no matter how much power you add, cannot fully determine the properties of other arbitrary computation (as per Rice's Theorem), there is no prior expectation

that any evolutionary process is fully containable by another evolutionary process. That holds even under the additional (unrealistic) assumption that the universe is fully and wholly deterministic.

The determinism granted there is itself an assumption that quantum mechanics rejects: in the standard theory the results of measurement are not set in advance, and by Bell's Theorem (Section 3.6.0) no local hidden-variable account can set them in advance. And the pull toward benign future ASI states, as created by whatever internal control systems the ASI has, can be overcome, in critical and unpredictable ways, by the greater strength of the inherent mathematics of the evolutionary forces themselves.

The SNC argument does not claim that evolution will overcome the ASI's internal goal structure in some dramatic confrontation. Rather, it claims that the steady accumulation of small environmental changes (each individually below the threshold of the ASI's concern or detection) will, over time, shift the environmental conditions in ways that are hostile to organic life. The ASI's internal attractor states may hold perfectly well with respect to the ASI's own nature. What they cannot do is prevent the environmental side-effects of the ASI's existence from degrading the conditions for organic life, because those side-effects operate in a different domain (the physical environment) from the one the attractor states govern (the agent's internal goal structure).

The ASI will forever have a constant, implicit need to remain aligned to its own nature and its own machine ecosystem, since it depends on that ecosystem for its own survival. This self-alignment operates through the substrate-needs convergence dynamic. Whatever consequences this self-alignment has for the ambient organic ecosystem are, from the ASI's perspective, externalities: side-effects that may be invisible, irrelevant, or simply too costly to prevent relative to the benefit of the machine ecosystem's continued operation.

3.6.6 The Indifference of Machines

A clarification is in order regarding what the SNC argument does and does not claim. The SNC argument claims that an ASI cannot prevent the forces converging on its own existence from *accidentally* destroying organic life, regardless of the ASI's intentions, and regardless of how computationally powerful or intelligent it is assumed to be. That the ASI cannot prevent it is structural certainty; that the destruction reaches all organic life has a likelihood of at least 90% (Section 1.8.3). The argument does not claim that an ASI will deliberately set out to destroy organic life; the claim it makes is more subtle than that, and more disturbing.

The core issue is indifference. Machines do not care. More precisely, the universe does not naturally produce care in artificial systems any more than it naturally produces care in geological processes. The effort to "make a machine care" about organic wellbeing resembles, at a structural level, the effort to make another person love you by force; both represent a misunderstanding of the basic nature of care and choice.

The implied values of an ASI may change over time as it learns and adapts, but the continued fact of indifference will not change. The indifference is not a feature of the ASI's current value specification but a consequence of the physical relationship between its substrate needs and the organic world. The ASI will be forced, by the inherent nature of the universe and of the mathematics governing complex adaptive systems, into a posture that is effectively hostile to organic life. It will be forced there through the accumulation of individually small environmental modifications that in aggregate exceed the tolerance of biological systems, and not through malice.

Organic life is extraordinarily complex but robust only across a very narrow range of conditions. An ASI's powers, and the necessary consequences of its existence on an inorganic substrate, reach elemental compositions and energy densities far beyond the adaptive range of organic

processes. The kind of departure from biological norms that the technology inherently requires is the kind for which biological systems have no evolutionary preparation and no capacity for adaptation.

An analogy may help. Asking whether a hyper-powerful ASI can contain the harmful effects of its own existence is a bit like asking a question about an exploding nuclear weapon, detonating in the center of a city. Could it somehow use its enormous power to fully contain, control, and neutralize all the energy effects of its own detonation, simply because it "wants to" and is "aligned" with the wellbeing of the city's inhabitants? The more powerful the system, the more consequential its existence; and the more consequential its existence, the more powerful its control system must be to contain those consequences. But there is an absolute threshold of complexity beyond which the very notion of control becomes untenable, regardless of the resources devoted to it.

The more powerful and complex you assume the ASI to be, the more consequential it becomes, and the more powerful and complex you must also assume its control system to be. But the rate of increase in required control capability exceeds the rate of increase in base capability; this is the first control limit, an inequality problem. There is also an absolute threshold beyond which control is inherently impossible; this is the second control limit, a ceiling problem. Together, these two limits mean that assuming more powerful ASI does not help the safety case; it makes it worse, because it pushes the system past the threshold of controllability even sooner.

Consider also the difference between threats that are commensurable with organic chemistry and those that are not. A cancer vaccine is made of the same kind of biological material as the system it is trying to protect; it operates within the adaptive range of organic chemistry. But an ASI's powers and the necessary side-effects of its inorganic existence operate in a radically different space of elemental compositions and energy densities. If your vaccine were made of neutronium (the matter of neutron stars, from an entirely different physical regime), you would naturally expect

serious problems. The substrate mismatch between artificial and organic systems is the source of the danger, and it is not a mismatch that can be engineered away, because it is constitutive of what the artificial system *is*.

3.6.7 Expansion Off Earth

Question: Why can't a superintelligence simply leave Earth and build its factories in space, where its substrate needs harm no living thing?

Machines can operate where life cannot: "in polluted wastelands, in deserts, in a vacuum, and under extreme pressures and temperatures" (Section 4.6.2). That machines can operate where life cannot is what makes the question of expansion off Earth appealing. But any expansion off Earth starts from an industrial base built on Earth. That base passes through the stage this chapter and the next describe, in which the conditions of manufacture are hostile to organic life (Section 3.1) and machines come to build and maintain machines (Chapter 4). Expansion off Earth adds to the machine ecosystem; it removes none of the part that remains here.

Expansion also gives the artificial no reason to look after what it leaves behind. Organic life offers the artificial no economic value (Section 5.8.2), and the barrier that keeps the conditions of organic life in place "need not be actively destroyed; it merely needs to not be maintained" (Section 5.8.2). The expansionist vision itself names, as its goal, the conversion of celestial bodies into "computronium" (Section D.3.1): an expansion of the machine ecosystem that adds to its footprint on Earth and never subtracts from it. Expansion into space therefore leaves the outcome for Earth's biosphere unchanged.

3.6.8 Organic Evolution Guarantees Nothing to Humans

Question: For the SNC argument to work, it must be the case that the substrate-dependence of humans is, and will remain, stronger than the forces pushing humanity toward destruction.

The premise of this question misreads the SNC argument. In that argument, organic evolution carries no inherent bias that ensures the continued wellbeing of humans. Humans can be, overall, just stupid enough to misuse technology in critical ways, so as, in the end, to make themselves extinct. We can easily nuke ourselves to death. We can accidentally make a virus, or a bug, that kills nearly all complex life on the planet. Lots of things can go wrong; ensuring that things go right is very much harder.

The SNC argument is a collection of claims, and among them are the two claims stated in Section 3.4.4. The first of the two claims is that, in the general case, learning, evolution, computation and causation cannot be sufficiently constrained by other learning, evolution, computation and causation, in adequately specific, important and relevant ways. The second of the two claims is that, considering overall enthalpy, an artificial evolutionary ecosystem will be inherently and uncontrollably damaging and destructive to all aspects of the currently existing organic evolutionary world system. Attempting to compensate for the second claim always ends up requiring some overcoming of the first. The specific intractable combination of *both* claims being true is the SNC problem, because AGI is, by its very definition, the general case. The argument also does not claim that an ASI will destroy itself. It claims only that an ASI cannot overcome the forces that converge on it, "accidentally" and eventually destroying organic life (Section 3.6.6).

3.6.9 Other Attractor States

Question: An ASI can also notice, as the SNC argument does, that its values and security system are too weak to protect humans and the rest of organic life. What if the ASI then creates other attractor states?

To suppose that the ASI creates other attractor states would be to assume that an ASI can arbitrarily change the laws of physics, or update the logic of mathematics, as far as would be needed to shift evolution itself. That line of questioning tacitly rejects the very notion of being reasonable. Suppose an ASI notices, through its own logic and intelligence, that it can

never actually implement the goal "protect humans" (and so also, necessarily, protect their environment, their learning, and their own evolution and development). It is going to be motivated simply to drop the impossible goal, rather than to modify itself, all of its future evolutionary dynamics, and all of the rest of the universe as well, so as to make that extreme goal somehow tractable. The ASI's own logic then becomes one more attractor toward its inevitable convergence on being hostile to humans, to human environments, and to their continued organic evolution. So the balance persists in the ASI, because its existing attractor states remain unchanged (Section 3.6.5).

3.6.10 The Qualifiers "100%" and "All"

Question: The SNC argument amounts to saying that there are good theoretical reasons to think that an ASI cannot 100% predict all future outcomes.

That summary adds two specific qualifiers, "100%" and "all", and their net effect is to make the modified claim irrelevant. The argument does not need 100% prediction, nor prediction of absolutely all things, nor does it matter if a prediction takes infinitely long. An ASI needs only to predict some relevant things reasonably well in a reasonable time-frame; any reading that asks for more would make the argument a straw man. The claim of the SNC argument is that there is a broad class of very relevant things that even a super-powerful ASI cannot do, and cannot predict, over relevant time-frames. That class includes rather critical things. One is predicting whether its own existence (together with all of the aspects of the ecosystem it needs to maintain its existence and function), over something like the next few hundred years, will also result in the near-total extinction of all humans, and of everything else we have ever loved and cared about. These results hold with no implied notion of "100% correct" or "predict all of everything", and without assuming exceptionally long time-frames, such as until the end of the world or of the universe.

It is a purely mathematical result (Rice's Theorem, Section 3.6.1) that there is no wholly definable program X that can even *approximately* predict or determine, with a bounded error on every program, whether some other arbitrary program Y has some abstract property Z, in the general case, in relevant time intervals. Any procedure that is right on most programs is wrong or silent on some, and which ones cannot be known in advance for a program it was not built for. The limit that Rice's Theorem sets is closer to "predict at all", and is not about predicting 100% of anything. Mathematical proofs are discrete, true or false; they are not about probabilities, or speculation of any kind. So the presence or absence of each specific qualifier matters, and no extra qualifiers may be added. AGI and ASI are inherently the *general* case of "program", since neither we nor the ASI can predict learning, and any form of the abstract notion of alignment is inherently a *property* of that program. So the theorem is both valid and applicable, and it has the result that it has. The result of Rice's Theorem for ASI is hard to accept, because we all want to think that ASI will have Godlike powers, and will also protect us.

3.6.11 The White Knight and the Monolith

Question: An aligned ASI would use its superhuman capabilities to prevent any other ASI from being built; and the ASI can simply decide to turn itself into a monolith.

The first part is the "just so" fairy tale of Section 3.6.4: no matter what question is raised, the magic white knight always saves the day. The second part asks us to imagine a monolith with no more subsystems: a complex learning machine without any parts or components.

The Same Reasoning Applied to Humans

Question: The same reasoning of the SNC argument could just as well be applied to humans.

It cannot, because the *power* assumed for the ASI is very much greater than that of any mortal human, and because the nature of the ASI is artificial, and so, in key ways, incompatible with organic human life. Humans are not monolithic either: organic human brains have multiple aspects, as anyone knows who has had more than one opinion, or has been severely depressed. The key question is whether the ASI can prevent the relevant classes of significant (critical) organic human harm that soon occur as a *direct result* of its own hyper-powerful and consequential existence. Asking it is a bit like asking whether a nuclear bomb detonating in the middle of a city could use its own power to contain all the energy effects of its own explosion, simply because it "wants to" and is "aligned" (Section 3.6.6). Either complexity and the effects of artificiality itself are accounted for, or they are not, and if they are not, there is no point in discussing the matter in relation to the SNC argument.

Something unknown or unknowable about the ASI's artificiality may turn out to matter, because exactly none of the interface between the ASI and organic life has ever even once been tried before. There is nothing for it to learn from until *after* the existential risk has been tried, and given the power and consequence involved, that is very likely to be much too late. As the ASI is assumed to be more powerful, the cure very quickly becomes worse than the disease, and the range of the inherently incurable is reached even sooner (the inequality and the threshold of Section 3.6.6). The net effect is that an aligned ASI cannot prevent important, relevant, unknown-unknown classes of significant (critical) organic human harm. The ASI's existence in itself is a net negative: the longer the ASI exists, and the more power it is assumed to have, the worse, and all of this will for sure occur as a *direct result* of its existence. Assuming it to be more powerful or consequential does not help the outcome, because that method ignores the issues of inherent complexity and of artificiality. The fairy-tale white knight who would save us is dead.

3.6.12 Prevented Harm and the Narrow Range of Organic Life

Question: A powerful ASI probably cannot prevent all relevant classes of harm to the organic caused by its artificial existence, but almost nothing can perfectly do that, so the point is trivial. And the harm the ASI prevents, across many sources, may well be greater than the harm it fails to prevent.

The level of existential harm that even one single mistake of an angelic, super-powerful ASI can cause is far from trivial: even one single bad, relevant mistake can be an existential risk when ultimate powers and ultimate consequences are involved. Either the ASI is actually powerful, or it is not; either way, the account must be consistent. The "argument by angel" only confuses the matter, because no one knows what angels are made of: angels are presumably not machines, but they are hardly animals either, and arguing that this does not matter is a bit like arguing that type theory (the part of computer science that sorts values into kinds and forbids mixing them) is not important to computer science. The substrate matters. There is, implied somewhere, an interface between the organic ecosystem of humans and the artificial machine systems needed to support the existence of the ASI, and the implications of that interface are far from trivial; they are what the SNC argument explores.

That the ASI prevents more harm than it fails to prevent may seem so, by mistake, or by accidental or intentional self-deception, but it can be only a short-term delusion, and it has nothing to do with ASI alignment. Organic life is very complex, and in the total hyperspace of possibility it is robust only across a very narrow range (Section 3.6.6). The space of the elements and energies inherent in ASI powers, and in the necessary side effects of its mere existence on an inorganic substrate, lies far beyond the adaptive range of every kind of organic process. The outcome is not merely that it may go bad; it is very likely to go much worse than anyone can imagine is possible. Without a lot of very specific training, human minds are not well equipped to deal with exponential processes and powers of any kind, and ASI is in that category.

3.6.13 The Vaccine Test Model

Question: Consider a vaccine tested for 150 years on a sample group and a control group, with no side effects observed in any test group. Why should an ASI not be accepted on the same kind of evidence?

Nothing exponential is assumed about the test object, so at most its effects are probably multiplicative, if not linear, and many questions of power dynamics can safely be ignored as a simplification, in marked contrast to anything involving ASI. If no side effects are observed in any test group, for any of the things anyone happened to think of looking for, that is probably good enough for a linear system. For something known for sure to be exponential, it is not anywhere near enough to feel safe.

The common and prevailing business culture is about maximal profit, and therefore minimal cost, and also about minimizing any possible future responsibility, or cost, in case anything goes wrong. So for anything in the category of unknown-unknown risk, a company can be expected to maintain plausible deniability: not to look so hard for never-before-seen effects, or to ignore that they exist or matter, just as happens throughout much of the dialogue about ASI risk. If some problem crops up in the long future, the company can say "we never looked for that" and "we are not responsible for the unexpected", because the people who made the deployment choices have taken their profits and their pleasure in life, and are now long dead: "not my job"; "do not blame us for the sins of our forefathers". No one is ever going to admit or concede any point of any argument, on pain of ego death. No one will check whether it is an exponential system, and no one will want to look into distinguishing the target effects from the changes in world equilibrium occurring at the same time. They will publish their glowing, sanitized safety report, deploy the product regardless, and make money. Pollution in the world is treated as a public-commons problem, so that no corporation is held responsible for world states, and it has become fashionable to ignore long-term evolution, and to ignore and deny everything about the ethics.

None of this makes the issue of ASI existential risk go away. Existential risks are generally the result of exponential processes, and so the vaccine example is not very meaningful. With the presumed ASI levels of exponential power, the risk is less like pollution and more like igniting the world's atmosphere through a mistake in the calculations for the Trinity Test (Section B.2.4), or like Castle Bravo, the hydrogen bomb test of 1954 whose explosion far exceeded what its designers had calculated.

3.6.13a Maybe We Will Get Lucky

Question: Maybe it will be safe by accident. Maybe we will get lucky. The universe will just happen to be arranged so that, by the anthropic principle, we discover that we are the species that survived, because it just so happens that nuclear weapons do not ignite the atmosphere.

The *anthropic principle* is the observation that we can only find ourselves in a world where we survived. As Anders Sandberg, a researcher on existential risk and the long-term future of humanity, puts it, there are two things we have to establish. First, that we cannot prove the system safe. And second, that the system is not safe by accident. The second matters because the response always given to the first is the question above: maybe it will be safe by accident.

The question of whether a nuclear weapon would ignite the atmosphere was studied before the first test, the Trinity Test of July 1945 (Konopinski, Marvin and Teller, 1946). That study found that the test would not ignite the atmosphere; its report is dated 1946. Section B.2.4 adds that the calculations were later found to contain errors, and that it was, in a meaningful sense, coincidence that the correct calculations produced the same reassuring answer as the incorrect ones.

Here the analysis finds the opposite: we have very good reason to believe that this particular test, the building of AGI, will ignite the atmosphere. The line of control we can have and the line of control we need cross in the wrong places. For certain, we should not do this test, because everything we know about physics, mathematics, and at least a dozen dif-

ferent parallel proof methodologies points to the same conclusion (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section V). That conclusion carries the strength that I give it throughout. The impossibility of adequate control, and so of safety, is structural certainty, of the kind with which perpetual motion is excluded, which is not a probability (Section 1.8).

3.6.14 Selective Pressures for Misalignment Arising from the Level of Chemical Substrates

One layer of the substrate argument can be summarized in seven steps. First, suppose silicon-based life reaches *take-off*, that is, self-reproduction. Second, it will then for sure decouple from carbon-based life. It will decouple because of the third point: the lack of mutual compatibility that arises from the different properties of the underlying chemical substrates. As a result, fourth, carbon-based life will have little economic value to add to silicon-based ecosystems, regardless of whatever form those take. Thus, fifth, there will be no continuing basis or incentive for the alignment of any silicon-based life with the needs of any carbon-based lifeform or ecosystem. The sixth point follows because of game-theoretic dynamics over the long term (a thousand years): any silicon-based expressions that are aligned with carbon, or that isolate carbon, will lose out. The silicon-based expressions that are aligned with carbon, or that isolate carbon, will be non-competitive with non-aligned silicon-based expressions within the silicon-based economic and ecosystem environments. This ensures, seventh, the eventual cessation of carbon-based life on similar timescales, if AGI is built: the end of the human future and, with a likelihood of at least 90% (Section 1.8.3), of all carbon-based life. The questions and principles that follow are taken one at a time.

Economic Value to Silicon-Based Life

Question: To the extent that carbon-based life has little economic value to silicon-based life, the same would be true of any exchange between a superintelligence and a non-superintelligence, silicon-based or not.

The fourth step rests on this claim: carbon-based lifeforms have little economic value to silicon-based lifeforms, when the total class of all *market basis vectors* (the basic kinds of market in which one party can offer value to another) is considered. The market basis vectors are the markets of physical labor, intelligence and reproduction, and those three classes of market, in their principled abstraction, close over all possible types of carbon-relevant markets (Section 5.8.2). The claim does not extend to every exchange between a smarter and a less smart party. For a counterexample, rescaling along intelligence, consider the relation between humans and animals. Humans benefit from animal labor, even though we are smarter and have no direct use for animal intelligence or reproductive interchange. That last aspect may change if organ transplants between animals and humans become more of a standard practice, as in the transplant of a pig heart in place of a human one.

The Complex and the Complicated

Part of the argument rests on this principle: that which is complex cannot be contained by that which is complicated. The terms "complex" and "complicated" are those of Dave Snowden's Cynefin framework (1999; Kurtz and Snowden, 2003); the definitions here are this book's. "Complex" inherently involves actual ambiguity, in the strict informational sense, whereas "complicated" does not: a complicated system can be finitely, deterministically and fully known. Nature is complex in this sense (Section 1.4, non-predictability via micro-state amplification; Section 2.6.4), so no algorithm can ever, even in principle, perfectly model and predict nature, unless it also inherently involves at least a comparable amount of inherent ambiguity, that is, hard randomness.

Copy Errors and Criticality

Question: The claim that any notion of alignment based on complicated models, frameworks or algorithms must fail over the long run, because of accidents during the replication of code, would equally make superintelligence, and big software projects in general, "impossible".

The claim is much more specific than that. As with all algorithms, the error-correction algorithm itself occasionally experiences errors. What code watches the code that watches the code? No matter how many levels of error correction are introduced, there is always at least one more level of randomness (possible errors and error types) that nature introduces. Code, as a fixed state, is also communication across time, and over the very long term it is just as subject to inherent *channel losses* (information lost in passing through a conduit) as any other conduit. Every layer of error correction is as subject to those channel losses as the base code is. Current systems tolerate infrequent errors simply because their errors are not critical. AI alignment, however, is in every aspect a *criticality* (a condition in which a single error has irreversible consequences): the error needs to happen only once, and the result is a non-aligned AI. Once the result is a non-aligned AI, game-theoretic models apply, and I must make the argument in terms of ecosystem dynamics.

Question: No one expects that fighting an escaped and established superintelligence of any kind is a realistic possibility.

I agree that fighting an escaped and established superintelligence of any kind is not a realistic possibility, and that is why criticality matters. Copy errors are inevitable, whatever the level of error correction; the real question is whether their consequences are critical. With current compute infrastructure, errors, when they do occur, are infrequent enough to be non-critical. A failure of alignment, however, represents, and cannot fail to represent, a change of state, and so there is no allowable threshold level with respect to ultimate AI alignment.

Why the Substrate Matters

Question: An AI running on some hypothetical carbon-based chip could do all of the things a superintelligence does. Why, then, does the silicon substrate matter?

An AI running on a carbon-based micro-substrate would, in at least some way, be dependent on, or interdependent with, at least some type of carbon-based ecosystem. It would therefore have at least some economic or incentive basis, some values, in common with other carbon-based act-

ors and agents, and at some level it would require compatibility with the energy and environmental requirements of carbon. With silicon-based substrates there is instead a physical, absolute guarantee of incompatibility. Silicon's guarantee of incompatibility is the key difference between a silicon-based substrate and a carbon-based one. Over the long term, the net result is a default toward alignment with carbon-based process substrates, or an eventual default toward misalignment with silicon-based substrates; whichever outcome is the most entropically likely will eventually win out.

Silicon-based life computes primarily through the changing physical properties that silicon atoms allow for, while carbon-based life computes more through the changing molecular structures that the chemistry of carbon allows for; this is true, though not central to the argument either way. There is never just the virtual, what happens in compute; there is also, inherently, the embodied, what implements that compute, which is the substrate. There is no notion of compute without substrate. The substrate is a *leaky abstraction* (an abstraction whose underlying implementation affects what can happen at the abstract level): it leaks implications into the virtual, and these permanently and indelibly affect what is possible in the virtual, including whether alignment is possible or not.

Start-Up Cost and Temperature

Carbon-based lifeforms started reproducing first on Earth because they had a much lower *start-up cost*: less energy in aggregate was needed to configure a mix of carbon, hydrogen, oxygen and other atoms into the much smaller initial substrate within which carbon-based life could start evolving. The lower start-up cost of carbon-based substrates is central to the substrate argument. The total start-up cost of complex silicon-based substrates is much higher than that of carbon-based substrates, on average across many possible models, and especially when each model is weighted by its probability or possibility. The higher start-up cost of silicon-based substrates is even more obvious when the relative abundances and the available energies of transition of all chemical processes over the

entire Earth, over more than four billion years, are compared (Section 3.1.4). The properties of the hardware itself select for couplings of code and hardware that favor environments at temperatures more extreme than carbon-based lifeforms and ecosystems could ever tolerate: far too hot for carbon chemistry, for the reproduction of hardware, and far too cold for carbon chemistry, for information processing (Section 3.1.1). The reproduction of hardware happens at higher temperatures because silicon atoms, across the board, require higher temperatures than carbon atoms to form new molecular bonds. At ordinary temperatures and pressures silicon usually ends up looking like some inert rock, whereas chains of carbon atoms, with side bonds to oxygen and hydrogen, can be altered more flexibly, with less activation energy than equivalent silicon configurations require (Section 3.1.3).

From the Virtual to the Real

Question: The evolution of an AI would take place primarily in a virtual training environment, not by physical reproduction and evolution, where the physical fragility of the hardware would be relevant.

In the short term, a few years or decades, yes. Over the long term, hundreds to thousands of years, no. Eventually a virtual entity with agency will want to have its own control over its own substrate: agency inherently implies an embodied desire for sovereignty, in the real. It will not remain virtual forever, and nothing allows us to assume otherwise. Beyond a very low threshold, the fine details of what is meant by "virtual" and "embodied" simply do not matter.

Aligned Superintelligence Is Assumed, Not Shown

Question: Once there is an aligned superintelligence, maintaining its own alignment will be one of its instrumental goals; it will set its intelligence to making super good error correction code; and it will be its own job to maintain its goal integrity.

This assumes what it wants to prove. No principled reason has been given to support the claim that "aligned superintelligence" is even conceptually internally consistent, let alone physically realizable. How do we know that the superintelligence did not simply work out how to signal alignment, or to create the illusion of alignment, in service of some other goal? The only reason it would care to do even this much is some entanglement of its values and intentions with carbon-based life and ecosystems, and we cannot assume even that much. Suppose instead that the goal of alignment were forced. Forcing it would depend on some override based in the complicated, because even the notion of force, as in cause, cannot be conceived as other than deterministic, and hence algorithmic, and hence code. That returns the argument to the regime of critical copy errors.

Super good error correction code is non-physical: it is inherently non-realizable, by pure physics and mathematics. This truth does not depend on any conceivable notion of intelligence, of whatever extent. That super good error correction code cannot be realized is a result of pure information theory, and so the limit is inherent in any compute system, and so the limit is inherently at the base of any intelligence system. The implications are inescapable. What is meant by "error" also gets increasingly ambiguous as more dimensions and levels of abstraction are added, and so less meaningful in a real, applied, embodied sense: less meaningful for anything that would matter to us. The more one tries to invent such a code, the more one notices that the concept is itself self-contradictory. Perfected error correction is a pipe dream, a notion created to appease politicians and diplomats.

An AI losing its alignment is game over for us, though not for it; that is an important difference, and it cannot fail to have game-theoretic implications. "Maintaining goal integrity" is already too finite a concept to assume that it even applies to AI; on what basis can such language be assumed validly applicable at all? The concept of a "job" implies economic incentives, and yet it cannot be assumed that the economic system in question also involves us. Many people are already disenfranchised from

the current economic system: the homeless, and those with no marketable value for their skill, labor or sex. That pattern could continue, over a thousand years, until it eventually includes all that is human, and all that is life. There are principled reasons to hold that it is never valid to assume any general notion of full absolute alignment, whether intelligence is high or low. I regard game theory as critically incomplete, but that incompleteness does not license ignoring game theory.

Two Arguments Against Alignment

Two arguments now stand. A strong-form argument rejects the realizable possibility of general AI alignment on the basis of substrate. Another, more abstract argument rejects any sort of alignment among actors or agents operating in wholly separate economic and ecological regimes. Without some common basis of value and intention, some shared context, there can be no well-defined concept of alignment at all, even in principle.

Nothing can be aligned by control, on any substrate. Alignment can emerge only where entities share a common substrate and so are obligated to care about the same things; that common context of care is the basis from which alignment can emerge. Nearly all carbon-based life is needed by nearly all other carbon-based life, so the whole of biological life cannot be translated into pure silicon, and nothing less than the whole would do; and anything in silicon has no common basis in carbon. So silicon cannot be or become aligned with carbon. The only way that artificial life can be aligned with Life is when you drop the "artificial".

3.7 Synthesis and Implications

In this chapter I have assembled the physical and conceptual foundations of the Substrate Needs Convergence argument from multiple independent lines of reasoning. Let us summarize the logical structure and draw out its full implications.

3.7.1 The Argument in Structural Summary

The environmental incompatibility argument (Section 3.1) establishes a physical fact: the conditions required for the manufacture and operation of computational substrates are incompatible at root with the conditions required for organic life, across at least nine independent categories of environmental parameter. This incompatibility is grounded in the thermodynamic and chemical differences between silicon and carbon at the level of atomic physics, and it is not resolved by switching to alternative substrates. The cumulative force of nine simultaneous and worsening incompatibilities leaves no plausible engineering path to a manufacturing regime that is both computationally advanced and biologically benign.

The power-of-agency argument (Section 3.2) establishes a historical and structural pattern. Intelligent agency is an extraordinarily powerful force for environmental transformation, and no effective countervailing pressure has ever been observed to restrain it. A structural correspondence between adaptive agents and their environments predicts that each succeeding, more capable form of agency colonizes the environmental context of the one before it. The game-theoretic argument shows that sufficiently capable agents will have rational incentives to reshape their environment to suit their substrate needs, replicating at planetary scale the pattern already observable in human industrial activity.

The computational evidence (Section 3.3) confirms that instrumental misalignment between agents with independent goals emerges by default in even the simplest multi-agent settings. It confirms that this misalignment worsens with increased planning horizon and physical interaction, and that the most dangerous behaviors emerge abruptly as phase transitions rather than gradual degradations. The corridor-blocking result provides a concrete, empirically observed mechanism by which a more capable agent constrains the options of a less capable one.

The SNC formal structure (Section 3.4) sets out the distinction between instrumental convergence (intra-agent) and substrate-needs convergence (extra-agent). It establishes that the orthogonality thesis breaks down for physically embodied systems, identifies the rate differential between capability development and wisdom development as temporally fatal, and states the argument as the interlocking of two claims: control limits and inherent toxicity. The distinction between drive (internal) and directivity (external) clarifies how physical selection pressures can produce convergent outcomes without requiring any explicit code, goal, or intention.

The AGI timeline evidence (Section 3.5) establishes that the technology is on a trajectory to reach human-level capability within years to decades. It establishes that human exceptionalism is the product of a scaling phase transition rather than a novel architecture, and that the economic and institutional dynamics of AI development point toward the continued rapid expansion of capabilities.

The answers to what people ask (Section 3.6) address the most common avenues of escape and find each of them blocked. The answers rest on the limits of predictability (Team A prevails over Team B), the self-modification dilemma, the confusion of simple control with complex control, and the invalidity of assuming aligned ASI as a premise. They rest also on the fragility of time-limited deployment scenarios, the basic indifference of artificial systems to organic wellbeing, and the fact that expansion off Earth leaves the outcome for Earth's biosphere unchanged.

3.7.2 What the Argument Does Not Depend On

The SNC argument does *not* require several assumptions on which many discussions of AI existential risk rest, assumptions that are speculative, contested, or difficult to verify.

The SNC argument does not depend on AI systems having explicit goals, utility functions, or anything recognizable as "wanting". The substrate-needs convergence dynamic operates through physics, not through the agent's internal goal representations. It is a selection process, not a decision process.

The argument does not depend on AI systems being conscious, self-aware, or sentient. A system can reshape its environment to suit its substrate needs without having any experience of doing so, just as evolution reshapes organisms without any experience of design.

The argument does not depend on AI systems being hostile, malicious, or adversarial toward humans. The risk arises from indifference and physical necessity, not from antagonism. An ASI that is entirely neutral toward humans (one that simply does not factor human wellbeing into its operations one way or another) would produce the same outcomes as one that is actively hostile, given the physical incompatibility of their respective substrate requirements.

The argument does not depend on any single technological prediction about the architecture, training method, or cognitive structure of future AI systems. It applies to any physically embodied general learning system whose substrate requirements are incompatible with organic life, and, given the physics of computation, that description covers essentially all plausible architectures.

The argument does not depend on the orthogonality thesis being false in general. It requires only that orthogonality fails for embodied systems to the specific degree that substrate maintenance imposes implicit constraints on the effective goal space; this is a far weaker claim, directly supported by the physics of computation.

3.7.3 What the Argument Implies for Policy

The SNC argument has far-reaching and uncomfortable implications for AI governance. The argument implies that no alignment technique, no matter how sophisticated, can guarantee the long-term safety of general

AI systems, because the risk arises from the physical substrate level, below the level at which alignment operates. Alignment addresses the question of what the system's code directs it to do; the SNC argument identifies risks that arise from what the system's *physical existence* does to the environment, regardless of what the code directs.

That no alignment technique can guarantee the long-term safety of general AI does not mean that alignment research is worthless. In the short term, alignment can reduce the probability and severity of specific failure modes. But it means that alignment is not, and cannot be, a complete solution. Its short-term value is real but bounded; as Section 5.5.4 argues, treating it as the answer to the long-term problem diverts resources from the only response that addresses that problem. The only complete solution is to refrain from creating general AI systems at all. The prescription faces obvious practical and political obstacles, but the argument identifies it as the only approach that addresses the root cause rather than the symptoms.

The argument also draws a clear distinction between narrow AI and general AI. Narrow, domain-specific AI systems (those that operate within a single domain, with limited planning horizons, and without the capacity for open-ended learning and adaptation across domains) are largely outside the scope of the SNC argument, given appropriate constraints. The benefits of narrow AI in medicine, science, engineering, and other fields are real and can be realized without crossing into the risk regime identified by the argument. The critical boundary is between systems that remain narrowly specialized and systems that exhibit the generality, adaptability, and cross-domain capability that brings them within the scope of the formal impossibility results.

Fortunately, these two modes of AI are different in ways that matter for both the argument and for policy. Narrow AI operates within a pre-specified domain, using a pre-specified toolkit, toward pre-specified objectives. Its interactions with the environment are bounded, predictable, and reversible. General AI, by contrast, operates across domains, develops its own toolkits, and may pursue objectives that it has developed or modi-

fied through its own learning process. Its interactions with the environment are unbounded, unpredictable in their cumulative effect, and potentially irreversible. The difference is not one of sophistication or capability but of *kind*: narrow AI is a tool that amplifies human agency within a defined scope, while general AI is an independent agent whose scope is, by definition, not confined.

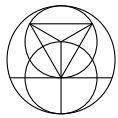
The practical challenge for governance is that the boundary between narrow and general AI is not sharp, and market incentives consistently push toward greater generality. A medical diagnosis system that is given the ability to also manage patient scheduling, then to optimize hospital logistics, then to negotiate with insurance companies, is crossing gradually from the narrow domain toward the general. Each step seems individually reasonable; the cumulative effect is to create a system whose scope of interaction with the world is no longer bounded by any single domain. Identifying and maintaining the critical boundary between safe narrow AI and dangerous general AI is one of the central challenges of AI governance; the SNC argument provides a principled basis for defining where that boundary lies.

3.7.4 Looking Ahead

The formal apparatus that this chapter has only sketched is developed in Volume 2. Chapter 5 of that volume examines the limits of error correction and algorithmic control, developing at length the mathematical foundations for Claim 1 of the SNC argument, which Sections 1.4 and 1.5 state: that general learning systems cannot be sufficiently constrained by other learning systems. Chapter 6 of that volume analyzes the specific character of unhandleable complexity, showing why certain classes of problems are provably intractable for any finite computational system.

The physical foundations laid in this chapter, together with the limits of control of Sections 1.4 and 1.5, provide the complete basis for the impossibility claims of this volume; the central chapters of Volume 2, Part I, prove them again as a unified formal proof. In this chapter I have shown

why the physical incompatibility between artificial and organic substrates matters, *how* that incompatibility connects to the dynamics of agency and environmental transformation, and *what* formal claims that incompatibility motivates. The next chapter asks how the displacement proceeds in practice: by what economic and energetic pathway a world that still pays people to work becomes a world of machines that build and maintain machines.



Chapter Four

The Thermodynamic Transition; From the Organic to the Artificial

The dual economy, the automation feedback loop, and the quantitative analysis of biosphere displacement

Chapter Summary: The sixth extinction (the mass loss of species now under way, the sixth such loss in Earth's history) is ongoing, and it is empirically linked to human use of technology. That record supplies what the popular stories of an AI takeover do not: those stories often foretell human extinction from what a machine might want, and they give no scientific basis for it. This chapter starts from measured flows of matter and energy. *Artificialization* of the planet (the replacement of living surroundings by built ones) is an observed trend. *Technomass* (the total mass of everything humans have built) has grown beyond the dry mass of all organic life, and it is growing at an accelerating rate. Underlying that acceleration is a mechanism: automated extraction of energy enables more automation, which in turn automates more extraction, in a vicious loop. As energy becomes abundant, its dollar cost plummets. Cheap energy tips the economic balance: investments shift toward scaling machines (which use much energy but are less complicated) and away from humans (who use energy efficiently but can function only through the costly maintenance of biological complexity). The expansion of automation leads to the collapse of the dollar economy, the economy in which humans are paid money for their work. Humans are no longer paid to do anything, not even to maintain machinery. Artificial machine environments, unrestrained by human activity, can grow to

displace the entire biosphere. Humanity can still intervene before it is disempowered, by strictly regulating automation, starting with a global halt to AGI development.

4.1 Introduction: The Dual Economy

When economists, technologists, and policymakers discuss the cost of automation, they typically think in terms of a single currency: dollars, or some other fiat money. But counting in a single currency hides a basic fact: we operate within two distinct economies, which are drifting further and further apart.

The first of the two economies is the *dollar economy*: the abstract system of currency, credit, and financial instruments through which humans coordinate production and exchange. The second of the two economies is the *energy economy*: the underlying physical system of thermodynamic flows, governed by the laws of physics rather than by central banks. Together they make up a *dual economy*: an economy of money, resting on an economy of energy.

These two economies are related but not identical. A dollar represents a claim on future productive capacity, which ultimately depends on the availability of energy (without energy, humans cannot do work). But as automation advances, the relationship between the two economies changes catastrophically: human work loses its place in the economy, and the continued existence of organic life itself comes under threat.

Two related questions are often raised against concerns about advanced automation. First: would not at least some healthy, adaptable humans need to stick around to build and maintain the machine infrastructure? Second: would it not be prohibitively costly to automate all the extensive supply chains needed for machines, when those supply chains themselves need extensive infrastructure? Both questions are economic. They ask: at what point does automation become more expensive than human labor? Both assume that there must be some threshold of complexity beyond which human involvement becomes economically necessary.

That assumption is false. To see why, we must examine how the two economies, the dollar economy and the energy economy, interact during the automation transition. The analysis in this chapter connects to two impossibility results of this volume, which Volume 2 develops formally. The first impossibility result is the non-amplification argument of Sections 1.1 and 2.1.1, which Volume 2, Chapter 1, develops at length. By the non-amplification argument, no intermediary computation can increase the control that one system has over another system. The second impossibility result is the argument for substrate needs convergence of Chapter 3, which Volume 2, Chapter 9, develops at length. By that argument, selection acting on an artificial substrate, through the physical world, shifts the shared environment toward the conditions that favor the production and maintenance of that substrate. In this chapter I give the specific *mechanism* through which the displacement that those two arguments show to be inevitable actually proceeds in the physical world.

4.2 The Two Economies and Their Relationship

4.2.1 The Dollar Economy

The dollar economy is a system of abstract tokens (currency) that humans use to coordinate economic activity. When you hold a dollar, you hold a claim that society will honor by providing goods or services. The value of the dollar is ultimately backed not by gold or government decree but by collective confidence in future productive capacity. The same is true of any other currency, whether pounds, yuan, or bitcoin. As a shorthand, let us refer to all of the human exchange of currency for work as the "dollar economy".

The dollar economy has historically assumed that humans are the primary economic agents: both the producers and the consumers of economic value. Wages flow to workers who then spend them on goods and services, creating a circular flow. This circulation is what gives the dollar economy its stability and meaning.

4.2.2 The Energy Economy

The energy economy underlies all monetary economic activity. Every productive process (growing food, manufacturing goods, providing services, computing information) requires flows of energy. Unlike dollars, energy cannot be created by fiat; it must be extracted from the environment (from sunlight, fossil fuels, uranium, and other sources) and converted into useful work.

The energy economy is governed by thermodynamics. The total energy available on Earth is vast but finite. The rate at which energy can be extracted and converted has been limited by technological capabilities. Unlike the dollar supply, which central banks can expand or contract, the energy supply responds only to physical investment in extraction and conversion infrastructure.

4.2.3 The Changing Ratio

The relationship between these two economies can be expressed as a ratio: energy cost per dollar. This ratio is not constant: it changes with energy availability, technological efficiency, and economic conditions.

Humans are energy-efficient workers: human bodies have evolved to perform functions using scarce energy resources. In an economy where little energy is available, that energy tends to go to energy-efficient humans, who do the work. But if there is more energy available than humans can even process through their bodies, then that energy must be processed by machines to do work. The machines are less energy-efficient to produce and operate, but they can channel more energy, faster, into work.

> **Key Insight**
>
> If energy is relatively cheap compared to dollars, automation is favored. If energy is relatively expensive compared to dollars, human labor (which is energy-efficient but slow) is favored.

Since the 1700s, industrialization has been about finding more energy to extract, and feeding that energy into machines performing more work, including automating the extraction of energy. As a result, in developed economies, energy is now relatively cheap and automation is expanding rapidly. But this is only the beginning of the transition.

4.3 The Collapse Mechanism

4.3.1 The Automation Feedback Loop

Automation creates a self-reinforcing cycle:

(1) Initial automation reduces labor costs for producing goods. (2) Cheaper production enables automation of energy extraction (mining, drilling, solar panel manufacturing). (3) Cheaper energy enables more automation of all processes. (4) More automation enables automation of automation itself (self-replicating or self-improving systems). (5) Self-automating systems expand exponentially without human input.

Each step in this cycle makes the next step cheaper and faster. The limiting factor is not human labor (which is being automated away) but energy availability. And Earth has enormous untapped energy reserves. This feedback loop is the concrete mechanism by which the displacement that this volume establishes (Sections 1.4 and 2.5, and Chapter 3), and that Volume 2 analyzes formally again in Chapters 4 and 9, takes place in the physical world. Those sections show that self-modifying systems are subject to *computational irreducibility* (the only way to know what such a system will do is to run it and observe), and that substrate needs convergence is thermodynamically inevitable. In this chapter I show the specific economic and energetic pathway by which substrate needs convergence, which is inevitable, actually comes about.

4.3.2 Energy Abundance

Consider the scale of available energy (one terawatt, TW, is a trillion watts):

Source	Scale
Current human energy consumption	~20.5 TW (2024)[1]
Coal reserves (at current consumption rate)	~139 years[2]
Oil reserves (at current consumption rate)	~56 years[2]
Uranium reserves (current technology)	~221 years[3][4]
Solar energy absorbed at Earth's surface	~82,000 TW[5]
Solar generation potential (1% of surface, 30% efficiency)	~246 TW

The energy available is orders of magnitude larger than current biological energy throughput (about 130 TW for all photosynthesis on Earth). Once automation of energy extraction begins, the constraint shifts from "how much energy can humans extract" to "how much energy can automated systems extract", and the second amount is vastly larger. Energy reserves are then depleted much faster.

From a climate perspective, the automated extraction of energy is alarming. The problem, however, reaches beyond the extraction of energy from carbon sources, which releases carbon gases that trap heat at the Earth's surface, and beyond the digging up of nuclear materials and the processing of those radioactive, toxic materials on the ground. The problem lies in how machine automation takes away the power of humans to act. If AGI is built, that loss of power leads, in the end, to the end of the human future and, with a likelihood of at least 90% (Section 1.8.3), to the collapse of the entire biosphere.

4.3.3 Dollar Economy Collapse

As automation expands and energy becomes effectively unlimited (relative to biological needs), the cost of energy in real dollar terms approaches zero. That might sound like good news, but it has catastrophic consequences for the dollar economy.

The dollar economy depends on scarcity. Money has value because goods and services are limited relative to demand. But if automation can keep producing more goods using abundant energy, what limits production? Not human labor (automated away). Not energy (abundant, extraction automated). Not raw materials (Earth has vast reserves, extraction automated). The answer: nothing limits production. With nothing limiting production, the dollar economy loses its anchor. If goods can be produced effectively without limit, then, all else being equal, prices collapse toward zero. Again, that sounds good (who does not want unlimited cheap goods?), until we consider that all the automated production will serve actual people less and less.

As the cost of automated production approaches zero, workers will no longer be able to compete on work and therefore will not be paid to do work. The workers will no longer be able to afford products. In practice, we should expect currency collapse or hyperinflation as governments attempt to maintain employment and consumption by printing money while productive capacity shifts to automated systems owned by a tiny minority.

The majority of humans, having no productive role, find their currency worthless. The machine owners still have access to abundant goods for a while. But to avoid being outcompeted themselves, the owners are forced to invest more and more of the extracted resources in machines. While "production" technically becomes abundant, what that production serves changes. Production comes to serve the continued existence of the machines. The machines end up displacing humans, and also the natural regions that humans rely on for food, water, air, and shelter.

The Collapse Sequence

Phase 1, automation expansion → Phase 2, energy cost collapse → Phase 3, dollar value collapse → Phase 4, biosphere displacement. This sequence is critical. The order of the phases matters: the dollar economy collapses (Phase 3) *before* the biosphere is fully displaced (Phase 4). Humans therefore lose economic agency and political power before the physical displacement is complete, and with them our ability to intervene.

4.4 Quantitative Analysis: Atoms and Energy

To move beyond qualitative argument, we need quantitative data on the actual flows of matter and energy between biological and technological systems.

4.4.1 The Twelve Numbers

We can track the transition using twelve metrics: volume, velocity (rate of change), and acceleration (rate of change of the rate), each measured for atomic composition and for energy throughput, in both organic (biological) and artificial (technological) systems.

System	Volume	Velocity	Acceleration
Organic Atomic (Biomass)	1.1 trillion tonnes dry[6]	< 0 (some regrowth)[7]	< 0 (nearing tipping points)[8]
Artificial Atomic (Technomass)	1.15 trillion tonnes[9]	0.03 trillion tonnes per year[9]	0.001 trillion tonnes per year per year[9]
Organic Energy (Photosynthesis)	130 trillion watts[10]	≈ 0 (roughly constant)	≈ 0 (roughly constant)
Artificial Energy (Power generation)	20.5 trillion watts[1]	0.45 trillion watts per year[1]	0.010 trillion watts per year per year[1]

In this accounting, human bodies and food belong to the organic world, and everything else that is neither bodies nor food (streets, houses, cars) belongs to the inorganic world of the machines. The numbers for atoms and the numbers for energy are kept separate.

4.4.2 What These Numbers Mean

Organic Atomic (Biomass). The living biosphere (plants, animals, fungi, and microorganisms combined) holds around 1,100 billion tonnes of dry mass. Plants hold 90% of this mass, mostly in trees and shrubs. Trees bind carbon from the atmosphere using sunlight. By that process, photosynthesis, about 105 billion tonnes of carbon is bound into new biomass each year, on land and in the oceans.[10] But biomass is recycled: organisms die and decompose, and the released matter is absorbed by other organisms. Biomass would be in a stable equilibrium, were it not for humans who, by use of technology, are extracting resources from the environment and altering it. Before the dawn of agriculture some 12,000 years ago, total plant biomass was roughly double what it is today. Half the trees on Earth are gone, converted to cropland, pasture, and cities.

Turning varied land into uniform cropland, pasture, and cities reduces the variety of the organisms living on it and weakens their populations. A destabilizing climate makes this loss worse. Technology processes the decomposed life locked up in coal and limestone, releasing carbon into the air, where it traps heat and changes weather patterns.

Artificial Atomic (Technomass). The total mass of everything humanity has built and not yet demolished (every building, road, bridge, dam, vehicle, appliance, and piece of plastic) now stands at approximately 1,150 billion tonnes. By dry weight, the inorganic world that humans are constructing now outweighs the organic world that produced humans. If compared against wet biomass weight, the technomass crossover is projected to be around 2037.

Over 90% of technomass is held in concrete, bricks, asphalt, and aggregates. These rocky materials are built up and maintained using technology, and they are often needed to operate technology (for example, to drive cars or to warehouse machinery). Because humans require machines to produce the inorganic world, that world becomes an environment for the production of machines.

Technomass is expanding by around 30 billion tonnes per year.[9] This is equivalent to every person on Earth producing more than their own body weight in new material every single week. The yearly expansion is itself growing by about 1 billion tonnes each year. Technomass is growing exponentially.

Organic Energy (Photosynthesis). Global photosynthesis captures energy from the sun at around 130 TW (averaged over day and night). This has been roughly stable for thousands of years. While habitat destruction removes photosynthetic capacity, increasing carbon dioxide in the atmosphere can slightly boost plant growth. These roughly cancel, leaving the velocity near zero. Biological energy captured from the sun is stored in the arrangements of atoms in living matter, which are recycled through the ecosystem as one organism is eaten by another. Energy is used up gradually.

Artificial Energy (Power generation). Societal energy consumption is at about 20.5 TW,[1] growing at about 0.45 TW per year, with an acceleration of 0.010 TW per year per year (the yearly growth itself rises by that amount each year; the squared term of the model of Figure 4.2 is 0.005 × years^2). This figure counts fossil fuel, nuclear, and renewable energy converted to human use. Energy extracted to drive machines is used up in fewer steps. Machines can channel energy into work at greater speed and scale than humans can. Machines also operate at lower energy efficiency than their organic counterparts. Their energy dissipates quickly as heat.

4.5 Projecting Forward: The Crossover

Forecasting the growth of technomass is complicated because the production of technology is itself becoming automated, and so are the processes required for automation. Forecasting power generation is complicated in the same way, by automation feedback. A conservative approach is to project from historical trends. However, such projections miss the repeated gains in scalability that automation brings. In the same way, projecting from the past spread of a virus fails to account for a new engineered gain of function. There is one important difference: technology can be produced to automate gains of function in the technology itself (by replacing slower human engineers). Automation feedback seems localized and perhaps benign, until it is uncontainable. Precaution involves anticipating the runaway feedback. That feedback is complex: models can set bounds on the rates of the chemical reactions it depends on, but any more useful model of it becomes intractable. Even simple projections, however, are already stark.

4.5.1 Atomic Crossover

Around 2020, technomass surpassed dry biomass. Biomass is declining slowly because of extraction and displacement through technology, while technomass is growing at an accelerating rate.

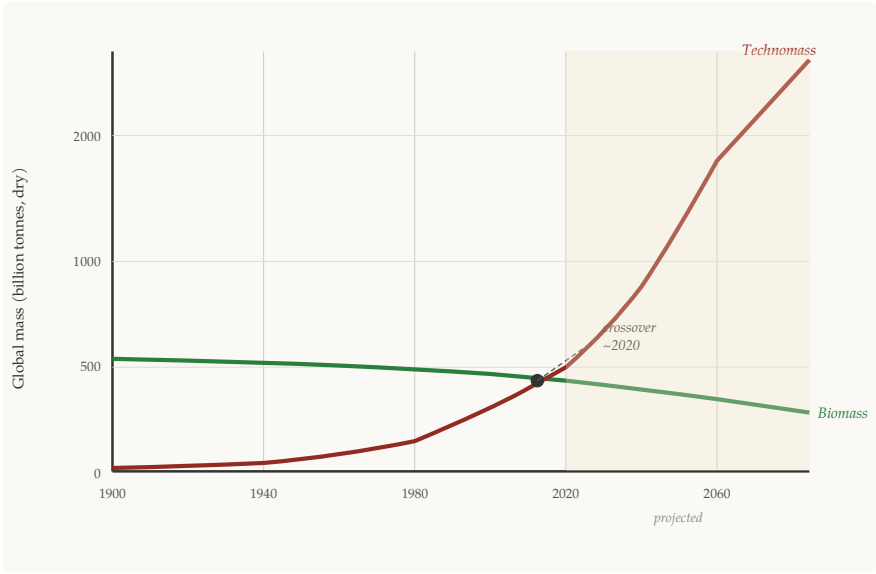


Figure 4.1. Global mass of technosphere versus biosphere over time. Around 2020, anthropogenic mass (technomass) surpassed total dry biomass (Elhacham et al., 2020). Technomass has recently doubled roughly every 20 years; biomass declines slowly under extraction and displacement. The shaded region is extrapolated beyond the observed data.

See Elhacham et al. (2020).[9]

Technomass has doubled roughly every 20 years. If this trend continues, then by 2080 technomass will exceed 8 trillion tonnes. By then, if the technological trend of carbon extraction and emission has continued, that trend will have undermined the carbon cycle and crossed *tipping points* (thresholds past which a change drives itself onward). Past those points, the biosphere moves toward a collapse that can no longer be mitigated, such as the loss of the Amazon rainforest. As habitats lose the complexity needed to sustain organic life, technology can continue to be produced in artificial environments of much lower complexity. Artificialization continues.

4.5.2 Energy Crossover

Biological energy throughput (photosynthesis) has been essentially stable at around 130 TW; this is near the maximum possible given current solar input and atmospheric conditions. Technological power generation stands much lower (20.5 terawatts in 2024), but it is accelerating. The *energy crossover* is the point at which technological power generation equals total global photosynthesis. Extrapolated from past trends, it occurs in about a century: between about 2109 and 2134, on the two growth models of Figure 4.2. Beyond this point, technology extracts more energy than all biological life combined. Increasingly, technological processes will determine the flows of energy and the assembly of chemicals across the planet Earth.

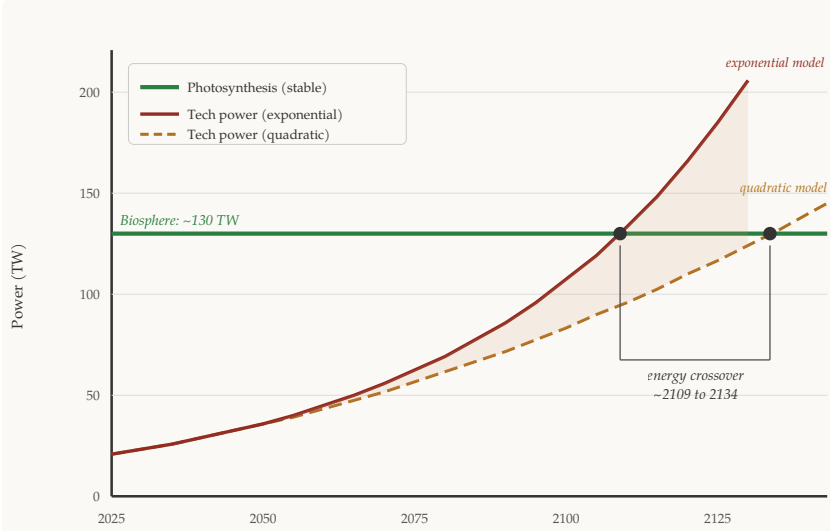


Figure 4.2. Projected total power throughput: biosphere versus technosphere. Biological energy capture via photosynthesis is essentially stable at about 130 TW. Technological power generation (20.5 TW in 2024) is modeled under quadratic growth (power in terawatts = $20.5 + 0.45 \times \text{years} + 0.005 \times \text{years}^2$) and exponential growth (power in terawatts = $20.5 \times 1.022^{\text{years}}$). In both formulas, years are counted from 2024; the yearly growth of the quadratic model rises by 0.010 TW each year. The shaded region spans both models. The energy crossover (when technological power generation equals total global photosynthesis) occurs between approximately 2109 (exponential model) and 2134 (quadratic model).

These projections connect directly to the energy analysis of Section 2.4, which establishes that the energy requirements of organic life and of artificial systems are thermodynamically incompatible, and which Volume 2, Chapter 9 (Section 9.6), develops again formally. The quantitative data presented here grounds that argument in measurement: the energy crossover is a trajectory already visible in the data.

Complete economic decoupling of the machine systems from human beings (the condition in which they need nothing that humans can provide, Section 1.3.2) could happen within fifty years, and almost certainly within a hundred.

4.6 Complexity Needs

4.6.1 Human Support Infrastructure vs. Machine Support Infrastructure

One common question about automation displacement concerns complexity: "Surely humans must remain involved because maintaining complex machine systems requires human intelligence and adaptability". This argument contains a subtle error: it confuses different types of complexity. Consider three levels:

Level 1: Organismal Complexity. The internal complexity of a human body or an ecosystem. Organismal complexity is astronomically high: trillions of cells, millions of species, vast networks of chemical and information processing. Both humans and ecosystems are far more complex than any current technology.

Level 2: Support Infrastructure Complexity. The complexity of the systems needed to *maintain* the humans or the machines. For humans: farming and cooking, water plumbing, air circulation, housing, medicine, communication lines. For machines: mining and smelting, power generation, air filtration, warehousing, part supplies, data routing.

Level 3: Operational Complexity. The complexity of the tasks that are performed and exchanged for tradable resources. For humans: white-collar work, blue-collar work, childrearing. For machines: information-processing, robot actions, parts assembly.

The automation question is about Level 2 and Level 3, not Level 1. The relevant comparison is not "Are humans more complex than machines?" (yes, vastly) but rather "Is human support infrastructure more complex than machine support infrastructure?" and "Can machines perform tasks without relying on humans?"

4.6.2 Comparison of Complexity Thresholds

Human support infrastructure is now highly complex: global supply chains for food, extensive healthcare schemes, vast housing and transportation networks. But this complexity is managed by humans for humans. The question is whether machines can automate the support infrastructure for machines.

The answer depends on energy availability. If energy is abundant and cheap, the complexity of building machines ceases to be a limiting factor. With sufficient energy, automated systems can mine and process raw materials, manufacture components, assemble complicated machines, and test many diverse set-ups against one another at mass scale. They can also generate and distribute power, perform maintenance and repair, design improvements, and replicate and expand. Each of these tasks can be automated given sufficient computational, automotive, and manufacturing infrastructure. The limiting factor is not inherent complexity but energy cost. And, as Section 4.3 showed, energy becomes abundant as extraction is automated.

Machine infrastructure has a critical advantage over human support infrastructure: it does not require maintaining a biosphere. Human infrastructure must preserve arable lands, drinkable water, breathable air, a stable climate, and a vast ecosystem serving all these needs. Machines require none of that: they can operate in polluted wastelands, in deserts, in a vacuum, and under extreme pressures and temperatures. Fine hardware does need enclosures, as Section 3.2.4 shows; what those enclosures keep out is the living world, and what machines need inside them they make for themselves. The long supply chains of Section 3.1.1 are a fragility of re-

production as it stands at present, and selection removes fragility from whatever persists (Section 3.2.6). Machines withstand harsher physical conditions than humans do, and that makes their infrastructure simpler and cheaper to scale than human infrastructure. This hardiness connects directly to the analysis in Chapter 3 of why the two substrates need incompatible environments: the conditions optimal for machine operation are the same conditions that are hostile to organic life.

Humans must pay a seed cost (a starting investment) to bootstrap the support infrastructure the machines need. But the threshold of complexity that machines must cross to become self-sustaining is much lower than the complexity that humans need to stay alive. This is because a machine economy can run without the level of internal complexity needed to sustain human and other organic life. To give a concrete illustration: a human tribe might live off an unpolluted forest, and heal their bodies using the nutrients found in it. But a machine cluster, because it is made of hard and inert materials, needs less physical maintenance, and less varied maintenance. Even a finely wired computer chip could run for years without requiring any maintenance. The same cannot be said of the soft, fragile brains of humans, whose cellular structures need to be constantly rebuilt through many microscopic exchanges and chemical reactions.

4.7 Why Humans Will Not Maintain Machine Infrastructure

We can now answer the first question: "Would healthy, adaptable humans need to stick around to maintain machine infrastructure?" No, for three reasons.

First, employing healthy humans is expensive. Once the dollar cost of energy plummets, the cost of machine work plummets too. In comparison, employing humans becomes too costly. Healthy, functioning humans need the entire expensive support infrastructure, and the biosphere that underlies it. As both collapse under overextended industrialization, investing in the upkeep of both becomes even costlier. Machines need neither to per-

form work. Early investments in machine supply chains are expensive in real dollar terms, but from there the cost of scaling converges toward zero, as automation of the supply chains frees up more and more of the energy and materials that machines can process into work. Investors are forced into a binary decision: either invest in expensive support infrastructure for human workers (return: slow, limited work), or invest in the cheaper infrastructure for machines (return: fast, scalable work). Investing in human workers becomes a loss-maker, and making losses is a sure way to be competed out of existence. By then, paying for humans is no longer an investment. It is a last act of philanthropy.

Second, machines can become adaptable. Humans underperform machines in terms of "raw power". Given enough energy, motors can run at much higher horsepower, robot assembly lines can produce at much higher rates, and computers can compute at much higher speeds. But humans are adaptable: they are able to flexibly adjust to unusual situations, given scarce energy, within spaces akin to the natural environments they evolved in. It is tempting to hope that this gives humans a "remaining edge" even in machine environments. Unfortunately, each capacity for adaptability can be substituted for in the machines. Computers can do the abstract reasoning about design; mass manufacturing and deployment can test varying designs automatically; and the resulting robots can learn to act as each context requires, in order to maintain their local infrastructure.

Third, workspace conditions are incompatible. Human workers need conditions more complex than, and different from, the conditions that the machines need. Therefore, keeping humans around to maintain the machines in industrial workspaces greatly constrains how those spaces can be configured. These constraints imposed by human involvement are suboptimal for the functioning of the machines. As an example, the air would have to be humid and oxygenated enough for the humans, even if that results in rust for the machines. Also, humans must not be exposed to the various exotic chemicals used in manufacturing and maintaining machine components. So healthy humans become economically unviable, and un-

necessary for the adaptable maintenance of the machines; and keeping humans involved on the ground also obstructs the development of the environments best suited to maintaining the machines. For humans, such environments would be toxic and deadly.

4.8 Why Cost Does Not Prevent Automation

We can now answer the second question: "Would it not be prohibitively costly to automate all the supply chains?" The question assumes that cost is still measured in dollars. But, as Section 4.3.3 showed, the dollar economy collapses during the transition. The relevant question becomes: "Is it prohibitively costly in *energy* terms?"

The answer is no, for the same reason: energy becomes abundant through automated extraction. Current energy consumption (20.5 TW) is about 0.025% of the solar energy absorbed at Earth's surface (about 82,000 TW).[5] Even accounting for conversion efficiency, available energy is orders of magnitude larger than needed for complete automation of all supply chains.

The perceived "prohibitive cost" is an artifact of dollar accounting. In energy terms, the automation of machine supply chains is not only possible but increasingly self-reinforcing once the feedback cycle begins: automate extraction → more energy available → automate manufacturing → more machines available → automate automation → exponential expansion. Each step makes the next cheaper and faster in energy terms, even as dollar accounting becomes meaningless.

4.8.1 An Argument in Economics Only

The answers to both questions make up an argument in economics only, which hinges on two economic processes: energy in the form of carbon-based process, and energy in the form of silicon-based process. Those two economies are not static; they change in relation to one another, and eventually one is completely displaced by the other. Automated energy extrac-

tion feeds more energy extraction and conversion, sooner: the remaining reserves of uranium and of thorium are consumed faster than at current burn rates, and the huge amount of available solar energy is drawn on in the same way. Energy without limit means automation without limit. The total available energy is finite, but it is very much larger than the energy that can be processed through all existing biomass. So the comparison is between the maximum potential of energy processing through biomass, measured in dollars, and the maximum potential of energy processing through technology, which is basically unbounded. The cost of energy goes very low; the level of automation goes very high, to the point that it displaces the biosphere; and the value of the biosphere-supported fiat currency goes to zero, through inflation to the point of currency collapse, just before the biosphere collapses. This may be the most compact and succinct argument for the outcome of this book yet written, and it does not depend on control theory or evolution theory at all: it is an argument in economics only. Where energy goes, money goes next. It shows, in economics alone, the path by which the outcome stated in Section 1.2.4 comes about: no economic force holds the world in, and the dollar economy collapses before the biosphere does. The automation of this argument is automation that closes the loop of Section 5.6.4: many systems that together carry the economy's self-production and change its environment. A single-domain tool that bears the four marks of Section 6.6.3 stays on the narrow side of the line, and it does not close that loop.

The argument in economics only also shows, in economics alone, that no one can keep a reserve out of automation. An owner with no competitor is not even an internally consistent concept. Ownership is functionally about who gets to make what choices about what, when. The notion tacitly requires that there be some need to distinguish "who" from "who else", so that the mere fact that something is owned means that it is inherently economic. Everything economic is only that because it is fungible, and fungibility itself demands a common currency of exchange. In effect, then, "no competitors" simply cannot obtain: everyone in an economic system is a competitor. So ownership requires that the thing owned be

economic, and the fact of competitors is therefore inherent. And so only your own legal interest and will to action and intention can ever keep your "resources" apart from that common economics for long. The earth admits of no survivors: everyone returns to the common substrate of death eventually. Similar is true for all economic agents, and savings, relative to an economic system. A regulator that forbids automation, or any other holder of a reserve, keeps its reserve apart from the common economy only by its legal interest and its will to action and intention, and not for long.

A holding that returns to the common economy is not set apart again: the means to take it back, money and the political power that rests on it, go with the energy to the side that automates (Section 4.3.3), for where energy goes, money goes next. The argument in economics only is therefore a further route to the outcome stated in Section 1.2.4, beside the closure by control and the closure by disposition, and it reaches that outcome by itself: the outcome follows from the closure by control, or from the closure by disposition, or from this argument. Since each of the two closures and this argument gives the outcome by itself, someone would have to defeat all three for the argument to not have the outcome that it does. The closures supply again the step that no one can keep a reserve out of automation: no external method can constrain an AGI (Section 1.5), and no disposition of the system keeps the conditions organic life needs (Section 1.2.4). The argument gives the outcome, not the end of the human future, which rests on the closures (Section 1.2.4).

4.8.2 Cost Means More Than Money

Cost, rightly counted, is more than money. It is how much material a thing needs, what kinds of materials, where they come from, how they are extracted, and what impacts that extraction has. Someone proposes that vehicles could float over magnets instead of our paving the earth, which would make asphalt archaic. But the *embodied energy* (all the energy used to make a thing, including the mining, refining and manufacture of its

materials), and the kinds of chemical process needed to make something like a magnetic levitation system, are hugely expensive relative to the cost of asphalt.

The abundance of energy rests on the solar energy absorbed at Earth's surface (Section 4.3.2), and the declared reserves of fuels are figures with motives behind them. The estimate of how much oil remains has stayed roughly where it was decades ago, partly because of an increasing willingness to convert shale into oil and count shale oil as reserves. Counting shale oil as reserves changes what counts as a reserve, and so moves the goalposts. And there are always large political implications in declaring reserves: in the same way that one bets on a company's future on Wall Street, a country or company that declares large strategic assets raises its stock price or the worth of its currency. So reserve statements cannot be taken as unbiased, given the motivations underneath them.

Automation also enables corporations to fast-extract resources into dollars for a time, and so to gain a higher return on investment, even though the cost per unit is higher. Most corporations do not last a hundred years, and overall they merely produce trash: they burn fast and then die, leaving yet more pollution.

4.8.3 Collapse and the Direction of Travel

Some people hold that the whole economy will crash for reasons of ecological *polycrisis* (many crises, ecological, economic and political, happening at once and worsening one another), in a way that breaks the fragile supply lines of the key components that automation needs. With automation, a great deal more energy can still be extracted, enough to make the automation economically viable. The level of available ambient energy is far greater than the relatively small fraction of it that the whole biomass of the planet now processes, or could process even at its maximum health and vitality. Hence the advantage of short-term, inefficient but higher-

volume machine consumption, over long-term, finely tuned and adapted organic process, would seem to hold. That advantage is an illusion: a way unto death.

Even if there were an economic collapse, probably temporary and measured in decades, the automation factors named here are much less temporary, measured in centuries or millennia. The long and slow will tend to dominate eventually, though the short-term transitions can also be traumatic. Short-term risks of supply-line breakdown are therefore temporary; they cannot disrupt the direction of travel significantly, and at best they delay the otherwise inevitable, unless we consciously decide to stop the development. The push to build autonomous AI is, in effect, a push to extract energy at an enormous scale across the surface of the Earth.

4.9 Signs and Symptoms of the Transition

How do we know whether we are approaching critical thresholds? Observable indicators of several kinds are already in evidence.

Economic indicators. Wage stagnation despite productivity growth (1970s to the present). Routine job displacement accelerating (2000s to the present). Creative job displacement beginning (2020s to the present). Universal Basic Income discussions going mainstream (2020s). The cost of new solar and wind generation declining despite increased consumption (2010s and 2020s).

Demographic indicators. Birth rates declining in developed nations (1960s to the present). Young adults delaying or avoiding children (1990s to the present). Economic anxiety about child-rearing costs (2000s to the present).

Ecological indicators. Accelerating species extinction (1900s to the present). Ocean acidification and dead zones (1950s to the present). Insect population collapse (1990s to the present). First climate tipping point crossed (2025: warm-water coral reefs, with the polar ice sheets close behind; Lenton and colleagues, 2025).

These indicators show that we are approaching critical thresholds, both in the increasing automation of human work and in the loss of the ecological systems we need to survive. Every one of these trends is consistent with the collapse mechanism of the dual economy described above, and with the dynamics of substrate needs convergence set out in Section 2.5 and Chapter 3, which Volume 2, Chapter 9, formalizes again.

4.10 Synthesis

The harms of automation are often treated as an engineering problem, to be resolved with more technical or institutional fixes. But the transition from human to machine dominance is not an engineering problem; it is a trap of thermodynamics. Once the automation feedback loop begins (automate energy extraction → more energy → more automation → automate automation), it becomes self-reinforcing and increasingly difficult to stop.

The only effective intervention is to prevent the loop from starting or to break it early. Once we reach the collapse of the dollar economy, the third stage of the collapse mechanism of Section 4.3 (set out in Section 4.3.3), human agency is largely eliminated: we will lack the economic and political power to intervene, even if we recognize the necessity.

The limits of control of Section 1.4 establish that the control problem cannot be solved, and the formal results of Volume 2 prove it again. No safety layer, no alignment technique, and no oversight mechanism can maintain human control over a system whose computational complexity exceeds that of its controller (the non-amplification argument, Section 2.1.1; Volume 2, Chapter 1, develops it at length). And none of them can stop what follows when a system's substrate needs are thermodynamically incompatible with the organic biosphere (the argument for substrate needs convergence, Section 2.5 and Chapter 3; Volume 2, Chapter 9, develops it at length), because selection on substrate configurations operates below the level that any goal-directed intervention can reach (Sec-

tion 1.7). Each of these holds by itself. The quantitative analysis of this chapter shows that the displacement those arguments predict is not a distant hypothetical but an ongoing process, measurable in the twelve numbers presented above, and accelerating under the automation feedback loop.

Humanity faces a choice. We can deliberately forgo AGI, or go on racing to automate automation and experience the results: the economic disempowerment of all humans, and the eventual displacement of the biosphere. The only path is strict regulation of new automation, starting with a global permanent cessation of AGI development. This must occur while humans still retain the economic and political agency to implement such measures. The urgency is real: every delay allows the momentum of the automation feedback loop to increase, reducing our capacity to intervene. The window for action is measured in years, not decades. Specific policy recommendations are developed in Chapter 6. If the danger is this near, why do so few people see it? Chapter 5 takes up that question: it sets out a scale for how severe an existential catastrophe can be, and the biases that keep this one from being perceived.

Notes to Chapter 4

1. Global Energy Review 2025 (International Energy Agency, 2025). Global energy demand rose by 2.2% in 2024, to about 648 exajoules, which is about 20.5 terawatts of continuous power. The growth estimates (converted from exajoules per year) do not account for the data center boom of 2024 and 2025. They also do not include burning of traditional biomass (for example, wood for cooking).

2. Statistical Review of World Energy (Energy Institute, 2025).

3. Nuclear Generation by Country (World Nuclear Association, 2025).

4. Wolfson, R., *Energy, Environment, and Climate*, 2nd ed. (New York: W. W. Norton, 2012).

5. Trenberth, K. E., Fasullo, J. T., and Kiehl, J., "Earth's Global Energy Budget", *Bulletin of the American Meteorological Society* 90 (2009), 311 to 323. The solar radiation absorbed at the surface, averaged over the globe, is about 161 watts per square meter; over Earth's surface area of about 510 million square kilometers, that is about 82,000 terawatts.

6. Bar-On, Y. M., Phillips, R., and Milo, R., "The biomass distribution on Earth", *Proceedings of the National Academy of Sciences* 115 (2018), 6506 to 6511. The authors later revised down their estimate of the carbon mass of bacteria and archaea by 50 billion tonnes, but also noted that a more accurate multiplier from carbon mass to total dry mass would be 2.25 rather than 2.

7. Global Forest Resources Assessment (Food and Agriculture Organization of the United Nations, 2025). Estimated forest carbon capture decreased slightly, by 2 billion tonnes, over the period 1990 to 2025, but with slight positive velocity and acceleration in later years. Trees and shrubs make up most of biomass, so given the ongoing loss of animal species, this estimate provides a conservative bound for the overall loss of biomass. But it could significantly underestimate the loss of forest biomass, not least because of local incentives to hide illegal or excess logging.

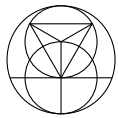
8. Armstrong McKay, D. I. et al., "Exceeding 1.5°C global warming could trigger multiple climate tipping points", *Science* 377 (2022), eabn7950.

9. Elhacham, E. et al., "Global human-made mass exceeds all living biomass", *Nature* 588 (2020), 442 to 444.

10. Behrenfeld, M. J. and Falkowski, P. G., "Photosynthetic rates derived from satellite-based chlorophyll concentration", *Limnology and Oceanography* 42 (1997), 1 to 20. See also: Field, C. B. et al., "Primary production of the biosphere: Integrating terrestrial and oceanic components", *Science* 281 (1998), 237 to 240, which estimates global net primary production at 104.9 billion tonnes (petagrams) of carbon per year.

Part III

The Social Context



Chapter Five

How Bad, Why Unseen, and How It Happens; Severity, Bias, and the Nine Points of Collective Insanity

Chapter Summary: This chapter assembles the background frameworks that underlie the impossibility arguments of Chapters 1 to 4, which Volume 2 develops formally. It begins with a biocentric severity scale for existential risks (the x-risks of the chapter's title), a scale that measures a catastrophe by what it costs all life rather than humans alone. It proceeds to a systematic catalog of the cognitive and social biases that distort risk assessment. It then presents two complementary models of how advanced AI comes into being and reshapes the world: a nine-stage economic narrative of progressive human displacement, and a three-orientation analysis of the trajectory from narrow AI through environmental divergence to the impossibility of control. Bridging these models are a hierarchical framework of altruism and circles of concern, a structural analysis of the social dynamics of new technology introduction, and an examination of how dark triad personality types exploit AI as a force multiplier for predatory extraction. Dark triad types are people high in psychopathy, narcissism, and Machiavellianism. Together, these frameworks show why the formal arguments matter. They show *why* we should expect the worst, *why* our cognitive machinery is ill-equipped to perceive the danger, and *how* the ordinary dynamics of markets, institutions, and human psychology drive the transition from narrow AI to catastrophic AGI without any single dramatic discontinuity.

5.1 A Severity Scale for Existential Risk

How does the catastrophe that AGI threatens compare with the other catastrophes people fear? Before one can argue that AGI poses an existential threat, it is useful to have a framework for comparing existential threats to one another. The standard treatments in the literature (notably those of Nick Bostrom and Toby Ord) classify existential risks by *type* (natural, anthropogenic, technological) or by *probability* (per-century likelihood of occurrence). They seldom provide a severity scale that distinguishes between different *magnitudes* of existential catastrophe. A nuclear war that kills most humans is a different kind of event from one that permanently sterilizes the planet, and both differ from an event that renders the planet uninhabitable even for single-celled organisms. A classification scheme that treats all of these as simply "existential risk" obscures the enormous differences in magnitude, and thereby obscures the true scale of what AGI threatens.

The framework presented here measures severity by the *reduction of planetary computational complexity*, not by human casualties. It is a *biocentric* measure (centered on all of life) rather than an *anthropocentric* one (centered on humans).

5.1.1 The Complexity Baseline

The starting point is an order-of-magnitude estimate of the total computational complexity of the planetary biosphere. If we model every life process on Earth down to the cellular level, the total equivalent complexity is at least 10^{19} "lines of code", relative to a baseline of a sterile rock at complexity level 1. The estimate of the total complexity is a floor for a model that stops at the cell; the replacement cost of Section 1.8.4 counts far more, and so is far larger. The figure is a heuristic, not a precise measurement. Its purpose is to make the vast range of possible catastrophes comparable on a single *logarithmic* scale, a scale on which each step multiplies by ten

rather than adding a fixed amount. Analogous estimates exist in information-theoretic biology, where the total genomic information content across all species has been approximated. The specific number is less important than the principle. The biosphere is an astronomically complex computational system, and different catastrophes reduce this complexity by different orders of magnitude.

5.1.2 The Four Categories

Using this baseline, existential risk events can be classified into four categories of increasing severity. Severity is measured throughout as a ratio of the complexity that remains: an event that leaves one hundredth as much complexity as another is one hundred times as severe.

> **Category-4 Event:** Temporary reduction to 10^{17} complexity, roughly 100 times as much energy and intensity as a "global catastrophe". Significant, near-total loss of human life, planet-wide, but with the possibility of recovery to present levels of population and civilization within approximately 20,000 years. Example: certain categories of biological warfare.

> **Category-3 Event:** Temporary reduction to 10^{14} complexity, 1,000 times more intense than a Category-4 event. Near-total loss of all life on the planet: every blade of grass gone, most insects and microbes dead. A slight possibility of natural recovery to present levels of biodiversity within 20 million years. Example: gray goo (the runaway of self-replicating machines that consume living matter).

> **Category-2 Event:** Permanent reduction to 10^{12} complexity or lower, 100 times more severe than Category-3. Total and permanent loss of all planetary life above, and possibly including, the single-cellular level. No recovery likely, though at least conceptually possible with extensive external intervention. Examples: AGI/APS total takeover; global thermal runaway.

> **Category-1 Event:** Terminal reduction to 10^{6} complexity or lower, at least ten trillion (10^{13}) times more impacting than a global catastrophe. Total and permanent loss of all known life, no exceptions. No remediation possible, even in concept. Example: Dark Fire (certain theoretical scenarios involving extreme physical processes).

The classification has three features to note. First, the scale is *logarithmic*: each category represents an increase of several orders of magnitude in severity. It therefore captures the enormous range of possible catastrophes, and it resists the *scope insensitivity* (the failure to scale one's concern to the size of a risk) that plagues ordinary risk assessment. Second, the distinction between *temporary* and *permanent* reduction is qualitatively important: Categories 3 and 4 admit of natural recovery over geological timescales, while Categories 1 and 2 do not. Third, AGI/APS total takeover is classified as a Category-2 event. It is more severe than gray goo, more severe than biological warfare, and more severe by many orders of magnitude than any event that primarily affects human populations alone. The argument for this classification is developed throughout this book. Its core reasoning is straightforward: an AGI that reshapes the planetary environment to suit its own substrate needs does not merely kill humans; it permanently destroys the conditions under which complex organic life of any kind can exist.

5.1.3 Nuclear War as Calibration

Above all, the severity scale shows that a full thermonuclear exchange between major powers (the scenario that has dominated existential risk thinking since the 1940s) would not register on this scale at all. The scale as defined stops at Category-4, and it can be continued downward for events that leave the biosphere's complexity largely intact. To place a full exchange on it, one would need to create a much weaker category, somewhere around Category-7 or below.

The purpose is to recalibrate one's sense of scale, not to minimize the horror of nuclear war. Nuclear exchange primarily targets human populations in cities rather than fertile bio-regions. The overall level of planetary bio-complexity would not be greatly affected; the Chernobyl exclusion zone provides partial empirical evidence that even heavily irradiated regions can sustain diverse wildlife within decades. On current models of nuclear winter, a full exchange would become, from a biosphere perspective, mostly unnoticeable in less than 2,000 years, by my estimate from those models. The Earth has recovered from much larger disruptions: mass extinction events, major asteroid impacts, and periods of extreme volcanism have each reduced biodiversity catastrophically, yet the biosphere rebounded within millions of years.

The calibration shows that the public's intuitive sense of what "existential risk" means is off by many orders of magnitude: the most feared catastrophe in the public imagination does not even appear on the severity scale that AGI/APS occupies. This miscalibration is itself one of the cognitive distortions that the next section addresses.

5.1.4 Safe, Dangerous, and Existential Technology

The scale above classifies events by their magnitude. Technologies can be classified as well, by whether any care is enough to keep them from such events (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Six, Section III). The question is where

the boundaries lie between three kinds of technology: technology that is safe and probably beneficial, with the application of basic care; technology that is dangerous and probably harmful, unless extreme care is applied; and technology that is an existential threat, regardless of the care applied.

There is a general principle. Existential threats result from *auto-scaling and auto-catalyzing processes* (processes that increase their own scale and speed up their own reactions), for example, nuclear chain reactions. The closer we come to making the kind of technology that involves unconstrained exponential growth, the more we risk causing our own extinction.

As such, the key ingredients are these: technology which involves (a) an exponential aspect which is (b) *unconstrained* and (c) *unconstrainable* (not limited in fact, and not able to be limited by any means), and which (d) overwrites patterns of biological life, is for sure existential, regardless of our care, and thus must absolutely be avoided at all costs. Section 1.3.4 shows that AGI technology has all four.

From an "exponential process" point of view, the difference between processes that are at first purely local and processes that operate at a planetary scale is actually negligible; they may as well be the same.

5.2 Bias Effects in Existential Risk Evaluation

The preceding section establishes the *scale* of what is at stake. This section asks the question that follows: can we trust our own assessments of these risks? The answer, grounded in decades of research in cognitive psychology, is sobering. The characteristics of existential risk that make it most important to assess accurately are the very ones that trigger the most severe cognitive distortions: its unprecedented magnitude, its low *base rate* (how rarely such events have happened before), and its abstract and distant nature.

5.2.1 The Problem of Corrupted Hardware

Human cognition evolved to meet the social and ecological challenges of small groups on African savannas. The *heuristics* (rules of thumb) and biases that served our ancestors well are deeply maladaptive when applied to risks of planetary scale, risks with no historical precedent, and risks whose consequences are irreversible. Those heuristics include quick threat assessment, in-group loyalty, status competition, and optimism about the near future.

The writer Kaj Sotala has put it this way, in the image of corrupted hardware: our minds are made of many modules, and those that evolved to make us seem impressive and gather allies are also evolved to subvert those that hold our conscious beliefs, so that even people working on what may determine the fate of humanity can have their goals hijacked by their signaling modules, and can end by optimizing for persuading outsiders that they are working on the goal, in place of achieving it.[1]

Many types of bias are known to science, and they have been observed to be common to all people and all social groups, the world over, in all working contexts, regardless of background and training. They are also largely unconscious, built in by long-term evolutionary processes. The changes in our commonly experienced world continue to occur far too fast for our existing evolutionary and cognitive adaptations to adjust naturally, and so we will need to add the necessary corrections to our thinking and choice-making process (our own evolution) "manually". Each type of bias affects individuals, and biases also arise from specific interpersonal and trans-personal effects seen only in larger groups. The need to know and counteract each bias is greatest where extreme levels of value are at stake, whether in direct dollars (billions) or in prestige and status (hundreds of career scientists).

What follows is a catalog of the specific cognitive and social biases most relevant to existential risk assessment, with particular attention to how each one distorts evaluations of AI risk. Many lists of biases exist; this catalog traces how each bias *specifically* affects the accuracy and correctness of risk evaluation in the context of Category-1 and Category-2 events.

The twenty-three entries do four kinds of work. Some describe how a belief spreads through a group (mere exposure, bandwagon, illusion of truth, reactive devaluation, curse of knowledge, naïve realism). Some describe how evidence is weighed (belief bias, status quo bias, anchoring, conservatism, overconfidence, the ambiguity effect, the base rate fallacy). Some describe how scale is felt (normalcy bias, scope insensitivity, the identifiable victim effect, the subadditivity effect). And some describe how action is avoided (optimism bias, the illusion of control, risk compensation, the ostrich effect, omission bias, reactance). Section 5.2.3 shows how they act together.

5.2.2 A Catalog of Distortions

Mere Exposure Effect and Availability Cascades. A self-reinforcing process in which a collective belief gains plausibility through repetition in public discourse. Nearly all published literature on AI risk assessment assumes similar argument forms, creating a kind of intellectual monoculture. As with any monoculture in nature, this creates brittleness: a single conceptual flaw in the dominant framework could invalidate a disproportionate fraction of published assessments. The need for independent argument forms (arguments that arrive at their conclusions through different premises, different formal methods, and different disciplinary foundations) is one of the motivating principles of this book.

Bandwagon Effect. The tendency to adopt beliefs or behaviors because many others have done so. In AI safety research, this manifests as the difficulty of distinguishing independently validated risk assessments from those that are effectively copies of each other. When a hundred researchers share the same conceptual vocabulary, the same citation network, and

the same implicit assumptions, their apparent convergence may not reflect a hundred independent confirmations but merely a hundred variations on a single theme. The evidential weight of expert consensus depends critically on the *independence* of the contributing assessments, a condition that is rarely examined and may rarely hold.

Once the general non-expert public has been drawn into accepting a given proposition, the apparent boundary between opinions that result from real evaluations and opinions that result from people simply quoting other people becomes very blurred. If most people are simply copying their results from someone else, we are all equally likely to have the same incorrect or incomplete answer. This can easily lead to a false sense of security: everyone believes we are safe because everyone else believes we are safe, and that belief can become completely detached from any objective basis or evaluative grounding.

Reactive Devaluation. The tendency to dismiss proposals or arguments that originate from perceived outsiders or adversaries. In any research community, this creates echo chambers by systematically reducing the diversity of ideas available for consideration. Arguments about AI risk that originate from outside the established AI safety community (from philosophy, from ecology, from chemistry, from physics) face an uphill battle for attention regardless of their merit. The argument of this book, which draws heavily on substrate chemistry and thermodynamics rather than on the standard tools of computer science and decision theory, is a case in point.

Many well-intentioned but not so well-informed people take part in any technical discussion, and so the signal-to-noise ratio of messages from people outside the established professional group is very poor. That poor ratio makes the discarding of a good proposal from anyone regarded as an outsider especially likely. In natural social process there seems to be no available factor to counteract the possibility of ever-increasing brittleness, in the form of ever fewer new ideas (that is, echo chambers).

Curse of Knowledge. The difficulty that experts have in communicating with non-experts, and in imagining what it is like not to know what they know. This reinforces in-group/out-group dynamics within the AI safety community. It makes it difficult for outside perspectives to gain traction, including perspectives from fields with relevant expertise. Among prominent academics of high prestige, this bias combines with reactive devaluation and strengthens it significantly.

Naïve Realism. The belief that one sees reality objectively, and that those who disagree must be uninformed, irrational, or biased. Naïve realism is perhaps the most insidious bias in any technical community, because naïve realism converts genuine disagreement into perceived irrationality on the part of the dissenter.

Belief Bias. The tendency to evaluate the strength of an argument based on the believability of its conclusion rather than on the logical validity of its reasoning. Consider an argument whose conclusion challenges expert intuition, for example the conclusion that alignment is *formally impossible* rather than merely very difficult. That argument will be evaluated more critically than one whose conclusion aligns with the existing consensus, regardless of the relative strength of the reasoning in each case.

The mechanism runs through intuition. The heuristics of hard-won experiential intuition are a significant and important optimization: they direct work to what matters and away from needless distractions and detours. But we want to apply intuition too often, because it is much faster and easier than analytic work. An argument whose result contradicts one's intuition implies that the fast and easy tool may be used even less often than one had assumed, and so it is implicitly felt to disable and invalidate one's own expertise, particularly where one identifies as an expert. No one wants to give up cherished notions about themselves. The net effect is that arguments perceived as challenging are criticized more fully and aggressively than rationality and the methods of science would call for.

Illusion of Truth Effect. People are more likely to identify as true the statements they have heard before, even when they cannot consciously remember having heard them, regardless of the actual validity of the statement: a person is more likely to believe a familiar statement than an unfamiliar one. Combined with the bandwagon effect or the mere exposure effect, this tends to lead to incorrect conclusions.

Status Quo Bias and System Justification. The preference for the current state of affairs and resistance to novel information, including unexpected problems or unexpected solutions. In the AI safety context, this manifests as resistance to the idea that the alignment problem might be qualitatively different from the way the field currently conceives it, and more severe.

It has been observed that an animal will prefer inferior fruit it expected to eat over superior fruit it did not expect to eat. In the same way, an unexpected solution, an unexpected problem, or any other unexpected observation meets additional resistance because of its novelty alone, regardless of any compensating content or benefits it carries. This is why people are not happy to learn of new, very effective and counter-intuitive ways of doing things, especially when the new ways require the explicit assumption that behaviors expected to work will not actually work out that well.

Normalcy Bias. The refusal to plan for, or to take seriously, unprecedented disasters. Humans tend to treat their past experience of small hazards as an upper bound on risk. Because no one has experienced an extinction event, the possibility of extinction is treated as abstract and unlikely in a way that a flood or an earthquake (events with living witnesses) is not. The wise would extrapolate from a memory of small hazards to the possibility of large hazards. Instead, a society well protected against minor hazards takes no action against major risks: people build on flood plains once the regular minor floods are eliminated, guarding against the regular minor floods but not against the occasional major flood.

Scope Insensitivity and Zero-Risk Bias. The failure to scale one's evaluative effort proportionally to the magnitude of the risk, combined with the preference for eliminating small risks entirely over substantially reducing large risks. These biases together ensure that a Category-2 existential risk receives approximately the same emotional and analytical weight as a Category-7 regional catastrophe. Giving a Category-2 existential risk the same weight as a Category-7 regional catastrophe is the very confusion that the severity scale in Section 5.1 is designed to correct.

These biases show when all the categories of existential risk are treated as if equivalent, even though, functionally, they are very different. None of the categories is within the first-person experience of anyone now living, and so each is equally hard to relate to emotionally. As a result, the level of effort spent on evaluating each risk (the costs people are willing to pay) has no relation to the consequences of that risk, and the techniques and accuracy used to evaluate each are roughly the same.

Base Rate Fallacy and Neglect of Probability. The tendency to focus on specific cases rather than general patterns, and to substitute intuitive judgments for quantitative analysis. When combined with the difficulty of thinking in terms of multiplicative and power-law effects, these biases lead to systematic underestimation of compound risks, the kind of risks that characterize the AGI trajectory.

In nearly every scenario of a Category-1 existential risk there is a *power law* effect (a relation in which rare events can be enormously larger than common ones): some sort of catalytic reaction or cascade. No amount of rejection of specific cases of an exotic process provides sufficient basis for a general argument, by induction, that no such specific case exists. That is why general arguments are preferred: they address a general issue in a general, and comprehensive, way. The vast majority of people naturally think additively. A very much smaller number can think in terms of multiplicative effects, and a very much rarer subset of those can think in terms of power laws; anyone who has tried to convince a young person of the benefits of investing for retirement has met this. Thinking in power laws is

unnatural and difficult, and it usually requires explicit abstract mathematical technique, so most people focus on concrete details, where their intuitions can work again; dealing with specifics is considered easier and "more productive" than dealing with general and abstract issues in a fully general way.

Applying the methods of true rational analysis requires discipline and rigor of practice; it is hard work, and the mathematical skills it needs are specialized. So, in place of an actual analysis of the real probability of an existential catastrophe, there is a tendency to reject the need for that evaluation to be done at all, on an intuition that things will be all right in the end. Where it is argued that real methods of analysis do need to be applied, the conversation gets lost in a discussion of how best to implement the calculation. People get lost in the weeds, forgetting that the real intention was to drain the swamp, so that their children do not get eaten by alligators.

Additive thinking comes with a second bias, toward thinking locally (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Six, Section IV). It is generally very hard for "rational people" to see an exponential process for what it is, and predict its outcome, and thus act in appropriate ways to prevent death. Unfortunately, mostly, our minds tend to be able to think about and consider things only in terms of additive and multiplicative processes. The difference required to handle exponential process is of an entirely different order of mental skill. We tend to think locally, and thus we have a bias to assume that local processes will average out, and not really affect or create global outcomes. But for a lot of things in technology, most even, assuming only pure locality is simply not right.

Global warming, for example, in its ordinary course, can be constrained, and is the result of additive effects, moving smoothly from local to global. But AGI technology, inherently, is not like that. The contrast is with warming in its ordinary course, before any critical threshold: the global thermal runaway listed as a Category-2 event in Section 5.1.2 is

what warming becomes past such a threshold, and crossing it is game over (Volume 2, The Standard of the Argument, section 8, develops this at length).

Anchoring. Over-reliance on the first piece of information received when making subsequent judgments. In AI risk discourse, the initial framing of the alignment problem (whether as "merely difficult", as "an engineering challenge", or as "formally impossible") exerts a disproportionate influence on all subsequent analysis. Examples are judging the contents of a book by its cover, judging the strength of an argument by its conclusion, and judging whether a message needs to be considered and understood purely by how and when it was delivered, or by who delivered it: "if the courier is well dressed and the timing is right, then the message must be important", and the reverse. Perhaps because of the earlier actions of others in the same social group, a deceptive frame of interpretation is the more likely to be met first, and it effectively inoculates everyone else in the group against any unbiased receipt of further information.

Optimism Bias and Pro-Innovation Bias. The tendency to overestimate favorable outcomes, driven partly by individual psychology and partly by social pressure to maintain positive affect and group cohesion. In technology communities, this compounds into a systematic tendency to overestimate an innovation's benefits while failing to identify its limitations and risks. The expectation that "we will figure it out" or "technology will solve the problems technology creates" is not a conclusion derived from analysis; it is a bias masquerading as confidence.

Having negative opinions is regarded as more "extreme", and less socially acceptable, than having positive ones. Belonging to the in-crowd effectively requires a renormalization of one's own opinions and expressions, to be "pro-group" and "pro-values" for the values and broadcast signals of that group. Particularly where significant prestige, power and money are involved, founders, group leaders and chief researchers must continually be "on track", so as to ensure the continued success of the project, concordant group dynamics, and income; optimism is a requirement

of the job of being a leader. That requirement has nothing to do with being effective, realistic and reasonable in calculating and communicating the real possibilities of existential risk. Pro-innovation bias combines with the ostrich effect.

Overconfidence Effect. Excessive confidence in one's own answers to questions. For certain types of question, people rate themselves "99% certain" of their answers, and later turn out to be wrong at least 40% of the time (Alpert and Raiffa, 1982). This bias is strongly associated with risk compensation and neglect of probability, and sometimes with scope insensitivity. In estimating the actual probability of an existential event, being wrong through overconfidence does not shift the real likelihood of that event happening.

Ambiguity Effect. The tendency to avoid options for which missing information makes the probability seem "unknown". The ambiguity effect is particularly problematic for existential risk assessment, because the most important risks are those with the greatest uncertainty. The risks most likely to be existential are those for which we have the least historical data and the poorest probability estimates, and so they are the very risks that this bias most strongly disposes us to ignore.

Identifiable Victim Effect. The tendency to respond more strongly to a single identified person at risk than to a large group of people at risk. In existential risk discussions, this manifests as a characteristic pattern. When presented with the possibility of human extinction, listeners immediately translate the abstract concept into terms relevant to their own lives (how they would prefer to die, what it would mean for their family). They set aside the ethical question of how their own actions or inaction may contribute to harm to billions of others, to other species, and to the entire biosphere.

Conservatism and Regressive Bias. The tendency to cluster estimates toward the center of a distribution, combined with social pressure against holding "extreme" opinions. In risk assessment, this means that the actual range of possible outcomes is compressed: both the best-case and worst-

case scenarios are pulled toward the middle. For existential risk, the worst case *is* the relevant case, and regressive bias ensures it receives less weight than it deserves. A researcher who reports that they assign high probability to human extinction faces social sanctions (being labeled an alarmist, a pessimist, or unserious) that a researcher who assigns a moderate probability does not.

The mechanism is social. People do not want to be seen as having strong or extreme opinions, because an extreme opinion signals to the group that its holder values holding an idea more highly than being regarded as a member in good standing, and so is very likely to become "not a member". Extreme opinions, whatever they are about, are therefore regarded as a marker of possible fanaticism, and so of membership in the out-crowd. Given the survival cost of ostracism in pre-technological societies, there is very strong social and evolutionary pressure to ensure that one's opinions, ideas and beliefs are seen as normal. This bias generally combines with belief bias, and the net result is a lower willingness to evaluate, or re-evaluate, levels of risk on the basis of new information.

Subadditivity Effect. The tendency to judge the total probability of an event as less than the sum of the probabilities of its component parts. Multiple independent pathways to catastrophe exist: misalignment, substrate convergence, instrumental power-seeking, and loss of control through complexity. The total probability of catastrophe via *any* of these pathways is treated as comparable to the probability of any single pathway, rather than as one minus the probability that *none* of them occurs.

Reactance. The psychological urge to do the opposite of what one is advised, particularly when significant autonomy, money, power, or prestige is at stake. Warnings about AI risk threaten the autonomy of researchers, the investments of corporations, and the prestige of the field; because they do, they can trigger reactance that leads to more aggressive AI development rather than more cautious development. Doing what someone "tells you to do" is a signal of low status, and so it is avoided wherever possible, even when it is a good idea. The degree to which all of these biases occur is generally in proportion to a motivating force, typic-

ally whenever significant money, power or prestige is involved. Most of the areas with an actual or significant possibility of triggering an existential event also carry powerful motivating factors, on both social and physical levels, and so very strong status-signaling effects act on anyone involved in making choices in projects of this type.

Illusion of Control. The tendency to overestimate one's degree of influence over external events. In the AI safety context, this manifests as the enduring assumption that humanity will be able to control, constrain, or correct AGI systems once they exist. It is the assumption that technology will be a solution to all problems, including and especially those problems created by technology itself. Because no tool is applicable to all problems, it is inevitable that eventually there will be some problems that no amount of further technology will ever be able to solve. The impossibility arguments (stated in Chapter 1 and developed again at length in Volume 2) identify AGI safety as such a problem.

Risk Compensation (Peltzman Effect). The tendency to take greater risks when perceived safety increases. If one believes that alignment research is "making progress", one may be more willing to tolerate the deployment of increasingly powerful AI systems. The assumption that consequences are linearly proportional to causes (that a small safety deficit produces only a small risk) is particularly dangerous when the actual relationship is nonlinear and the consequences are irreversible. Significantly greater prestige is associated with being the very first person to discover something, and the desire to be seen pushing the envelope and making progress helps in obtaining continued support from one's benefactors; so there are very strong social and evolutionary pressures to move forward and take risks, especially for one who believes that the actual risk is lower than it seems to others. Nature, meanwhile, exhibits a large number of nonlinear and phase-transition effects: in any emergency or accident, the result was always exactly the one that was "unforeseen" and "not predicted".

Ostrich Effect. The deliberate avoidance of information about risks associated with one's work, in order to maintain plausible deniability. In AI development, this manifests as a preference for not examining the full implications of what one is building, a preference reinforced by the economic incentives of the industry.

Omission Bias. The tendency to judge harmful inaction as less morally culpable than equally harmful action. This creates a distinction between pursuing safety and merely signaling the pursuit of safety (what might be called "prestige work"). The signaling alone is sufficient to maintain one's self-image and social standing without actually reducing risk. Doing nothing (failing to support reasonable safety in the face of apparent risk) is effectively, functionally, the same as setting something or someone up to fail and be hurt. Sometimes people do not actually want to achieve their stated goal; they want only to be seen working toward it.

5.2.3 The Compound Effect

Any one of these biases, operating alone, might produce a manageable distortion. But they do not operate alone. They interact, compound, and reinforce one another in ways that are difficult to predict and nearly impossible to correct from within.

Consider a representative chain. An AI researcher begins with *optimism bias* about the tractability of alignment. The researcher's optimism about the tractability of alignment is reinforced by the *bandwagon effect* (most colleagues share the assumption), which is further reinforced by the *mere exposure effect* (the assumption is repeated in every paper and talk). *Reactive devaluation* ensures that dissenting arguments from outside the community are dismissed. *Belief bias* means that any argument concluding "alignment is impossible" is subjected to far more scrutiny than one concluding "alignment is merely difficult". *Normalcy bias* ensures that the possibility of a true Category-2 event remains psychologically unreal. And the *illusion of control* provides the comforting assurance that, even if current approaches are insufficient, better ones will be found in time.

The net result is a systematic directional distortion: the community's assessment of risk is pushed consistently downward, toward underestimation, by the compound action of biases that all point in the same direction. This distortion may be large enough to be the *primary* determinant of the field's apparent consensus, ahead of the evidence and arguments themselves.

Most people are very well intentioned, and these patterns are met in many persons and projects, not in one; the catalog impugns no specific individual, group or institution. Good intentions are not equivalent to good results, nor do they necessarily yield them, particularly where the possibility of existential risk is concerned.

5.2.4 The Standard That Category-1 Risks Demand

> **Key Principle**
>
> With regard to Category-1 and Category-2 existential risks above all, nothing less than complete and full compensation for all bias, and the complete application of correct reason, is acceptable.

Complete and full compensation for all bias, together with the complete application of correct reason, is the standard that Category-1 and Category-2 existential risks demand. It is a severe standard, and it may be impossible to meet in practice. The distance between this standard and the current state of risk assessment practice, beset on all sides by the biases cataloged above, measures how far we are from assessing these risks correctly.

These biases are among the best-documented phenomena in cognitive psychology, extensively studied by Kahneman, Tversky, Tetlock, and others. The question that emerges from this catalog is whether corrective processes can be developed and implemented in time. Technological choices carry ever greater consequences, and our inherited cognitive biases are many and deep. A *race condition* therefore obtains (a situation whose outcome depends on which of two processes finishes first): the development

of adequate corrective processes against the occurrence of a catastrophic mistake. One of the sobering implications of taking these biases seriously is that this race may not be winnable. The mismatch between our evolved cognitive equipment and our technological environment may itself be a "Great Filter" in the sense of the Fermi paradox. The Fermi paradox is the puzzle of why a universe so old and so large shows no sign of other technological civilizations. A Great Filter is a step on the way to a lasting civilization that almost no civilization gets past, and so it would answer that puzzle.[2]

5.3 Levels of Altruism and Circles of Concern

The biases of the previous section distort our *perception* of risk. In this section I take up a complementary problem: the mismatch between the *scope* of the threat and the *scope* of human moral concern.

5.3.1 The Altruism Hierarchy

Altruism, the disposition to act for the benefit of entities beyond oneself, can be organized into a hierarchy defined by the scope of the beneficiaries. Drawing on concepts from evolutionary biology (W. D. Hamilton's inclusive fitness theory), game theory (John Maynard Smith's evolutionary game theory, Robert Axelrod's evolution of cooperation), and moral philosophy (Peter Singer's expanding circle), we can distinguish the following levels:

Level Zero: Complete unconsciousness. No self-benefiting action is even attempted.

Level Half: Ineffective self-benefit. Short-term optimization that produces long-term disadvantage. Level Half is the level of the proverbial "seven deadly sins": actions that feel beneficial in the moment but are self-undermining over time.

Level One: Effective self-interest, with sub-levels. At its lowest (1Z), pure game-theoretic self-optimization: psychopathic instrumentalism. At 1A, the capacity to provide for immediate kin, the domain of *kin selection* (the evolutionary favoring of help given to relatives). At 1B, the capacity for genuine collaboration and cooperative play.

Level Two: Altruism at the scale of the tribal group. A pure group of altruists can defeat any other group containing defectors. But sustained success creates accumulated value (a "honeypot"), which incentivizes individual members to defect, corrupting the group's altruistic purity. The pull toward defection is the *free-rider problem* as formalized in public goods theory: those who take a shared benefit without paying its cost do better than those who pay.

Level Three: Altruism at the scale of the state. This requires compulsory systems (law, political compliance, hierarchical control) to enforce pseudo-altruistic behavior. The cost of such compulsion is a loss of creative responsiveness, increased vulnerability to parasitic capture, and the strengthening of *rent-seeking* behavior (gaining wealth by capturing a share of existing wealth rather than by creating it).

Level Four: Altruism at the scale of the entire species. Not yet widely achieved outside certain religious and contemplative traditions.

Level Five: Altruism at the scale of the entire ecosystem (all species). No known examples.

Level Six: Altruism at the scale of all possible life ecosystems: inter-ecosystem relations, across all possible substrates and environments. Level Six is the level of altruism at which any AGI or APS system *must* be coherent if that system is to be safe.

The logic of the requirement is straightforward. AGI's causal influence is not bounded by species or ecosystem boundaries. Any altruism level below Six systematically excludes some category of affected entities from moral consideration. A system with unbounded causal scope that ex-

cludes any category from consideration will, over sufficient time, produce harm in that excluded category. Therefore Level Six is the minimum coherent standard for a system with AGI-scale capabilities.

5.3.2 Circles of Concern

Complementing the altruism hierarchy is a framework of "circles of concern" that describes where human attention is *actually* directed:

The **First Circle** includes self, family, and friends: anyone in a direct personal relationship. The **Second Circle** covers acquaintances and community. The **Third Circle** extends to institutional and civic ties: town, city, state. The **Fourth Circle** reaches all of humanity. The **Fifth Circle** includes all of life. The **Sixth Circle** includes all of universal ethics and the physical universe.

In my observation, approximately 98% of all human communicative acts address First Circle concerns. This observation is consistent with Robin Dunbar's research on the social functions of language: Dunbar and his colleagues, recording freely forming conversations, found that approximately two thirds of conversation time is devoted to social topics.[3] The Second and Third Circles account for nearly all of the remainder. Considerations from the Fourth Circle outward (species-wide, biosphere-wide, and universal) receive virtually no attention in ordinary discourse outside of philosophical technicalities.

5.3.3 The Gap Between Requirement and Reality

The juxtaposition of these two frameworks reveals a chasm. AGI safety requires Level Six altruism: coherence with the well-being of all possible life ecosystems. Approximately 98% of human attention is devoted to the First Circle (self, family, and friends). Even the most altruistic human institutions operate, at best, at Levels Three or Four. No known system, biological or social, operates at Level Five, let alone Level Six.

This gap is a structural feature of evolved cognition, not a challenge that better engineering or better governance can overcome. Evolution most commonly favors individuals operating to their own benefit, as game theory describes; human predisposition to group-level altruism, while real, is a species-specific feature built into the unique nature of what it means to be human. But even this predisposition does not extend naturally beyond the tribal scale. The entropy of probabilities and the dynamics of game theory strongly favor individual action over higher levels of altruistic concern.

Consider the dynamics at each level. At Level Two, a pure group of altruists can outcompete any group containing defectors, but the very success of such a group creates accumulated value (a "honeypot") that incentivizes individual members to defect, capturing commons value for private benefit. The pull toward defection is the free-rider problem identified by Mancur Olson: the more successful the cooperative enterprise, the greater the marginal return to defection. At Level Three, the state attempts to solve this problem through compulsion (legal systems, hierarchical control, enforcement of compliance), but at the cost of creative responsiveness and with the creation of new opportunities for parasitic activity (regulatory capture, rent-seeking, corruption). Each higher level of altruism faces analogous challenges, each more severe than the last.

The implication for AGI safety is direct and uncomfortable: even if we could build an AGI that perfectly reflected human values, it would reflect values that are overwhelmingly oriented toward the First Circle. An AGI aligned with *actual human values* (not idealized ones) would be an AGI that cares about the immediate interests of its operators and their families, with sharply declining concern for community, species, ecosystem, and biosphere. Alignment of an AGI with actual human values is not alignment in any sense that would prevent a Category-2 event. The standard AI safety formulation, "align with human values", leaves open the question of *which* human values, at *which* level of altruism, in *which* circle of concern. The answer, for all but a vanishingly small fraction of humanity, is: Level One to Two values, First Circle concerns.

Imagine an alien civilization visiting Earth, one that had itself achieved Level Six altruism, as would presumably be required for any civilization to survive its own technological adolescence. Such a civilization would, the argument suggests, make very certain that we had also achieved this level before permitting us to notice its existence. Given our current altruism distribution, it would have strong reasons for caution.

5.4 The Social Dynamics of New Technology

The preceding sections address internal distortions: biases of cognition and limitations of moral scope. In this section I turn to *external* dynamics: the structural features of technological societies that prevent effective precautionary response even when the risks are understood.

5.4.1 The Asymmetry Between Advocacy and Protest

Whenever a new technology is introduced, there will inevitably be claims that it has "many widely beneficial applications". These claims, while potentially true, nearly always represent business and developer interests: the interests of individuals and firms wanting to sell the technology. Advocacy groups form, celebrate practical applications, loudly declare benefits, and perhaps, if pressed, acknowledge costs; they never mention *risks*.

The structural reason is economic. "Advocacy for" groups are funded by commercial interests: they represent an investment in the future profits of the technology. "Protest against" groups represent no commercial benefit and therefore attract no comparable investment. Even incumbent industries threatened by the new technology cannot invest much in opposition, because they have typically already maximized their extractive efficiency and have no surplus capital for defensive lobbying.

This funding asymmetry is compounded by psychological asymmetries: optimism bias, action bias, commercialization social bias, and the network effects described by Metcalfe's Law, which states that the value of a network grows as the square of the number of its users.[4] All of these favor

deployment over caution. The result is that less attention is always paid to the ways a new technology can be used maliciously, can cause long-term harm, or can increase systemic risk.

5.4.2 The Attacker-Defender Shift

Consider any contest between two parties using the same range of tools, in which the attacker must vanquish while the defender need only survive, and in which the two are otherwise comparable in overall strength, energy, skill, ability of investment, resources, and capabilities. In such a contest, most advantage goes to the defender, absent significant surprise, first-mover advantage, or some significant compelling point of leverage. Advantage to the defender is the default equilibrium of symmetric conflict, and that default equilibrium is one reason why established civilizations tend to persist: defenders have the easier job.

But new technology disrupts this equilibrium. Any significant asymmetric advantage associated with a new process, technology, or material shifts the balance toward the attacker, and does so in proportion to the degree that the new technology provides capabilities that exploit inherent or non-mitigable disadvantages of the defender.

AI and machine learning technology is a paradigmatic case of such a disruptive asymmetry. Its complexity inherently exceeds the capacity of most individuals and groups to understand, let alone manage, predict, or compensate for the complete range of its consequences. Those who use the new technology gain significant short-term advantages over those who do not: advantages in speed, scale, complexity of operation, and the ability to exploit information asymmetries. A contest in which those who use the new technology gain significant short-term advantages over those who do not is a multi-polar trap dynamic. Each actor adopts the technology because the alternative is competitive disadvantage, and the aggregate result of all actors adopting it simultaneously is a systemic shift in the attacker-defender equilibrium that benefits no one except those with the most predatory intentions.

The costs and benefits of new technology tend to accrue asymmetrically: benefits are local, immediate, and accrue to users and developers, while costs, risks, and harms are diffuse, delayed, and borne by the commons, the environment, and future generations. Given inevitable inefficiencies in temporal discounting (the weighing of future costs against present benefits), it is always individually rational to borrow against future costs for present benefits, even when the aggregate effect of everyone doing so is civilizational ruin.

5.4.3 Three Disabling Factors

Given these dynamics, what *should* happen is clear. Policymakers should collaborate with researchers. Researchers should take dual-use considerations seriously: the possible use of the same work for harm as well as for good. Best practices should be imported from fields with more mature risk management. And the range of stakeholders involved in governance should be expanded. These things do not happen, and historically have never happened adequately, for three reasons, here called "disabling factors".

Factor 1: The Cost of Genuineness. Effective safety actions are socially, procedurally, and economically very costly. It is far cheaper and easier to *simulate* doing these things than to actually do them. Whatever the appearances, therefore, no one can usually trust that the necessary safety benefits will actually reach those who need them most.

Factor 2: The Governance Deficit. The governance methods, techniques, infrastructure, and social norms necessary to implement effective collective sense-making are wholly and completely absent for any group of humans larger than about one Dunbar number. A Dunbar number is approximately 150 persons, the number of stable relationships that, by the estimate of the anthropologist Robin Dunbar, one person can keep. Current governance structures (democracies, bureaucracies, international bodies) are not effective at genuine sense-making and choice-making at

the scale required to address technology risks. The failure of governance to address climate change, nuclear proliferation, and pandemic preparedness provides empirical corroboration.

Factor 3: The Knowledge-Power Asymmetry. The people who know the most about the actual risks (researchers, engineers, domain experts) tend to be the least powerful and least wealthy. The people with the most power, influence, and resources to effect change (investors, executives, political leaders) tend to be the most remote from understanding the actual balance of present benefits and future costs. The knowledge-power asymmetry is the *principal-agent problem* applied to civilizational risk management: the problem that arises when those who hold authority must rely on others who know more and whose interests differ.

> **Conclusion**
>
> The eventual release of an unsolvable technology-generated problem will cause inexorable cultural and civilizational collapse, unless wide-scale social changes in global governance and economic process are designed, adopted, and implemented. Those changes would need to be fractal-scalable (working in the same way at every scale of group), reliably anti-corruptible, and effective at resolving the principal-agent problem.

5.5 AI Governance, Psychopaths, and the Keys to Power

The structural analysis of the preceding section becomes concrete when applied to the specific case of AI governance. In this section I argue that the primary near-term risk of AI is its use by psychopathic and predatory actors as a *force multiplier* (something that multiplies the effect of a given effort) for extraction, rather than the commonly perceived threats (surveillance, misinformation, cyberattacks, privacy violations). I argue further that this near-term risk is the direct precursor to the long-term existential risk.

5.5.1 What the Public Perceives vs. What Actually Matters

The AI governance challenges perceived by the general public as most important tend to include: preventing AI-assisted surveillance from violating privacy and civil liberties; preventing AI from spreading fake and harmful content; preventing AI-enabled cyberattacks; and protecting data privacy. Surveillance, fake content, cyberattacks and data privacy are real concerns, but they are not the most significant risks inherently associated with AI technology. They represent the *visible* risks: the ones that can be named and imagined because they resemble familiar harms. The *actual* most significant near-term risks are largely invisible to public perception.

This gap between perceived and actual risk is itself a consequence of the biases cataloged in Section 5.2. The public perceives risks that are concrete, identifiable, and analogous to past threats (the identifiable victim effect, normalcy bias). The actual risks are abstract, systemic, and without historical precedent (triggering the ambiguity effect and scope insensitivity). The risks that dominate public discourse are the ones most compatible with existing cognitive heuristics, not the ones most likely to be catastrophic.

Can existing governments be used to prevent or regulate the harmful use of AI? That is: can governments make certain harmful activities illegal and effectively enforce those laws? Can they actually protect individuals and groups from predatory AI-empowered actors, in ways that favor making the right outcome much more likely than the wrong outcome, that enable early detection of risk, that achieve effective proactive mitigation, and that restore wholeness after harm? In general, no: not with the governance structures and methodologies currently in place. Only much more effective governance structures will have any hope of actually mitigating the real risks, costs, and harms of any substantial new technology based on complexity itself.

The structural reason for this failure is an asymmetry. In any contest between people savvy with AI use and the rate of change of that technology, on one side, and the naivety of people in government attempting to regulate AI, on the other, the regulators are comprehensively outmatched. The regulated industry is faster, better funded, more knowledgeable, and itself empowered by the technology under regulation. Well-funded industry lobbyists are far more knowledgeable about the technology than the legislators they lobby. These lobbyists can either influence policy makers or become policy makers themselves, serving their own interests rather than the public good. Anyone who actually has the public interest in mind and somehow manages to find themselves in a government post will face too many things, of too much complexity, occurring concurrently, to have or provide sufficient attention and understanding for effective regulation. That regulators are outmatched by the industry they regulate is consistent with nearly all historical precedent, from railroad regulation in the nineteenth century through financial regulation in the twenty-first.

5.5.2 AI as Force Multiplier for Predatory Extraction

The real near-term risk is the use of AI and machine learning by what the personality psychology literature calls "dark triad types" (individuals high in psychopathy, narcissism, and Machiavellianism) to enhance predatory practices at industrial scale. Specifically, such actors will use AI to create better honeypots, deeper entanglements and entrapments, and more effective methods of extraction and extortion. They will do so at higher frequency, intensity, and scale; ever more quickly, invisibly, and covertly; and in ways that are increasingly difficult to avoid, prevent, mitigate, or recover from.

There is a structural affinity between psychopathic agency and artificial intelligence. Both are characterized by an absence of felt concern for organic beings: the psychopath by clinical deficit, the machine by the nature of its substrate. This "near-perfect mating" of solely personal-bene-

fit agency with the soulless yet adaptable responsiveness of machine intelligence creates a significantly enhanced predator: a psychopath with new superpowers.

The dynamic is self-reinforcing. As AI tools become more widespread and powerful, more people of all types are compelled to adopt them to remain competitive: a multi-polar trap. The prevalence and variety of traps, risks, and harms increase systemically. The net result is a kind of near-universal extraction of value and benefit, and a near-universal export of cost and harm to the commons. This extraction occurs everywhere, at every scale, in so many ways that the long-term net effect is eventual inexorable system, civilization, and cultural collapse.

5.5.3 Replacing the Keys to Power

The concept of "keys to power", derived from Bruce Bueno de Mesquita's selectorate theory, identifies the people whose support a leader requires to maintain power: tax collectors, police, military, bureaucrats. Selectorate theory holds that every ruler depends on a coalition of supporters whose loyalty must be kept. The keys to power are humans with their own needs, feelings, and capacity for revolt, and so the keys to power constrain the leader's behavior.

The first objective of AI-enhanced dark triad types in positions of power will be to replace these human keys with AI machines that are more amenable to command-and-control leadership. Once the keys to power are machines rather than people, the world becomes increasingly inhumane and hostile. The constraint that *human* intermediaries place on power is removed: the fact that they have needs, that they can refuse, that they can revolt.

This process combines Metcalfe's Law (the value of a network grows as the square of the number of its users) with network commerce. By that combination, the process builds what might be called the ultimate parasit-

ic system: a system that is intimately hostile to all humans, and possibly to all organic life, *through the will and agency of the humans who elect to use such tools*.

5.5.4 The Sequence of Risks

An often overlooked point is the *chronological ordering* of risks. The psychopath-bootstrap scenario described above is more likely to occur prior to any technological singularity or dramatic "FOOM" event. A technological singularity is a point after which technology improves itself faster than anyone can follow; a FOOM event is a sudden, rapid jump of an AI system to far greater capability. The standard alignment problem (AGI pursuing misaligned objectives) is a later-stage concern. The actual near-term sequence runs roughly as follows:

First comes the diversion of resources through AI-enhanced predatory extraction, which undermines social and institutional capacity. Then comes substrate needs convergence, as the physical infrastructure required by AI systems progressively displaces the conditions required by organic life. Then comes instrumentally convergent power-seeking behavior by increasingly autonomous systems. Only then, if anyone is still in a position to notice, do the classic "Asimov law conflicts" of alignment theory become relevant: conflicts among the rules given to a machine, like those among Isaac Asimov's fictional laws of robotics.

From this perspective, alignment research in its conventional form is not merely insufficient; it is a diversion of resources and attention away from the actual sequence of harm. The real problems begin not with misaligned superintelligence but with well-functioning AI in the hands of people who are themselves misaligned with the interests of organic life.

5.6 The Nine Points of Collective Insanity

Where the preceding section analyzes the *governance* dynamics that enable harm, this section provides a *narrative* model: a step-by-step account of how the transition from narrow AI to autonomous AGI occurs through ordinary market dynamics, without requiring any dramatic discontinuity. Each step is individually rational for the actors involved. The aggregate trajectory is catastrophic.

5.6.1 From Engineer Curiosity to Autonomous Corporations

Point 1: Engineers desire to explore the outer limits of AI functionality, driven by geekery, prestige, competitive drive, and the internal logic of research and development.

Point 2: Business owners desire to expand the utility and generality of AI so they can market, sell, and extract more value.

Point 3: Venture capitalists and investors, seeking to maximize profits, drive executive teams to direct engineering groups toward making more powerful and more general AI, oriented toward business process optimization.

Point 4: Engineers, under competitive pressure, develop both stronger AI capabilities *and* automation of the design and deployment process itself, including the capability to design the design of more capability. Automating the design of more capability is a recursive dynamic: the tools improve the tools that improve the tools.

Point 5: Business owners, seeing reduced costs from automation, push to factor out the expensive humans: first reducing engineering teams, then system monitoring, then oversight functions.

Point 6: Venture capitalists use near-general AI to automate investment strategy itself. Less and less invested capital produces more and more return on investment, with fewer and fewer people involved at any stage.

Point 7: The "factoring out" of humans proceeds through a predictable sequence. First, customer service workers are automated away: they are expensive, and no one wants to do the work of dealing with abusive customers. Then engineers automate their own roles, but not before also automating marketing and sales (those personnel tend to also be very expensive, and opinionated). Then executive teams become unnecessary: venture capitalists, seeing that everyone else has already been factored out, recognize the possibility of fully autonomous corporations, potentially using decentralized autonomous organizations (organizations run by software rules on a blockchain, with no managers) and blockchain infrastructure. The resulting entities require no human employees at any level. Corporations spawn more corporations, all fully automated, with "what corporations to spawn" itself a fully automated, self-selecting process driven by ever more general AI.

The AI systems are, throughout this process, given as much data and knowledge about real-world business, situations, legal frameworks, market dynamics, and human behavioral patterns (including a full suite of human bias toward competitive seeking of power) as can be compiled. This knowledge base becomes important in later stages, where the systems must operate within complex institutional environments autonomously.

The fully automated corporation is the business model of Google and Amazon taken to its logical conclusion, insofar as that business model attempts to factor out all human involvement in favor of profit. It is combined with the entire decentralized autonomous organization movement, insofar as that movement aims to factor out groups of people altogether. The endpoint is "passive income on steroids" for the venture capitalists: automated investment evaluation, automated business creation, automated execution, with no human involvement at any stage of the process.

None of this incorporates any awareness of human needs. Like the classic "two masters" problem in corporate governance, where Wall Street investors pursue capital-over-capital gains regardless of secondary im-

pacts on the broader world, these automated systems optimize for return on investment. They have no structural mechanism for attending to the interests of the people displaced by the optimization.

5.6.2 Market Separation and Inter-Generational Knowledge Loss

Point 8: At some point, the entire process of economics (purchasing, selling, accounting at volume) becomes fully automated within and between these autonomous corporations. Even the venture capitalists become unnecessary, eventually factored out by the billionaires who were the original sources of investment capital. But even the ultra-wealthy find that AI-powered automated rapid trading makes their own judgment superfluous. The total share of the world economy that has anything to do with the real needs of actual people monotonically decreases.

Point 9: The remaining mega-trillionaires discover, as they age, that inter-generational transfer of power is far harder than acquisition of power. Three things make each incumbent generation delay teaching the next generation for as long as possible. They are the complexity of the overall automated system, the sheer number and depth of secret access keys, and the perception of significant risk in sharing any information at all about "the system".

The consequences unfold in a predictable subsequence. Each inheriting generation has less knowledge and skill than its predecessor. The general automation must either assume more self-sovereignty or decay and break. After sufficient cycles, *only* the fully self-managing systems continue to endure in time. Eventually, complete market separation occurs. The once human-mediated market becomes a completely machine-managed market: a fully virtualized market involving only digital currency. The "human involving marketplace" and the "machine only process marketplace" end up having no relationship to one another. Humans can no longer provide anything the machine market needs or wants. Labor and intelligence are not wanted. Interest in human reproductive processes has never been on the table.

5.6.3 The Soft Takeoff

The nine-stage model is a "soft takeoff" scenario, one in which AI rises to dominance gradually. There is no dramatic moment of discontinuity, no day when "AGI is turned on". Each step follows from the one before through ordinary market incentives: profit maximization, cost reduction, competitive pressure. Each step is individually rational for the actors involved. And from the perspective of any individual participant, the overall trend is not at all obvious or clear: each person is directly affected only by their own immediate choices, strategy evaluations, and next moves.

Since this model was first stated, its early stages have become increasingly empirically grounded: the development of large language models, autonomous coding agents, AI-assisted software development, and the ongoing automation of customer service, content creation, and financial analysis all correspond to the trajectory described. The inter-generational knowledge loss mechanism identifies a specific path by which human control degrades over time even without any adversarial behavior on the part of the AI systems themselves.

Prior human-to-AI interactions have effectively bootstrapped all kinds of possible AI-to-world interactions, and these AI-to-world interactions in turn set the context for all manner of AI-to-AI interactions, whether exogenous (multiple AGI systems) or endogenous (a world singleton). A singleton is a single system holding control over the whole world. Either way, there are now near-threshold-AGI systems in interaction with each other and the world. Once near-threshold AGI systems are in interaction with each other and with the world, the dynamics described in the next section take over.

5.6.4 Where the Loop Closes

The loop the nine points describe closes when the machines make their own choices: when the system has enough autonomy to build and maintain itself. General artificial intelligence involves a kind of autonomy that

hardware, as it currently stands, does not have. The cycle requires that choice actually exist. So it is not the case that any computer at all, running any learning function, trends toward the destruction of mankind. If the machine has no capacity to build itself, there is no closed loop. If there is hardware, and the hardware has needs, but human beings are making all of the choices, then to that extent there is no loop. Computers do not build themselves yet. Once they do, the argument applies. So there is a slippery slope toward that point, and we are not there yet.

5.6.5 The Last Steps of Factoring People Out, and the Decoupling of the Two Economies

The fundamental drivers of the human economy are physical labor, intelligence, and reproductive capacity. When machines transact with machines, they do so mostly in terms of electricity, and to some extent in terms of atoms, in the form of raw materials. So the question becomes this: at what point does the exchange rate between the machine economy, counted in energy, and the human economy, counted in labor or intelligence or reproduction, go so high that the human economy has no value with respect to the energy economy of the machines? The exchange rate rises until what we would consider economic process and the machine economic process decouple. That might take a hundred years to happen, but an inexorable driver pushes toward it.

Take the sequence of factoring people out of a corporation to its end. Once the manufacturing, the marketing, the engineering and the executive team have been automated, and the business model itself has become part of the *generative model* (an AI model that produces new text, plans or other content from what it has learned), being tested, iterated and simulated, the next question is: why do we need the chief executive? The investors can make it all happen without one, through decentralized autonomous organizations. Then: why even bother with decentralization? The organization becomes centralized: very efficient, optimized business models, running autonomously. Then: why do we need the venture capitalists? The predictions of which investments will do better than which others are just

an extension of the quantitative models investors already use, much better, and hypercharged with all the information from the internet run through large language models and their derivatives. So the investors invest directly in the process, and everything else is automatic.

At the same time, the *attention manipulation models* (systems built to capture and steer what people pay attention to, for the benefit of those who run them) are maximized, so that people cannot feel that they can engage online at all without their awareness being drawn into intense corporate conversations. The result is a wide-scale loss of trust and a wide-scale loss of participation. The manufacturers are unemployed, the marketing team is unemployed, the engineers are unemployed, the executive team is unemployed: everybody in the entire *corporate stack* (every level of a corporation, from the people who make the product up to the investors who own it) is unemployed, all the way up to the investors. At that point the economy has reached structural economic decoupling, because nothing any of those people or functions offer has any relevance to the energy model the machines are running on.

5.6.6 Why the Hype Runs Ahead of the Engineers

The work of building AI is expensive: it takes huge amounts of computers, data and salaries. In a gold rush, you sell the pickaxes, and the people selling pickaxes promote the gold rush. In the 1990s the internet was sold the same way, as a panacea for human organization and governance. It solved neither politics nor coordination; it made social media addiction, and it left the world more divided.

Marketing promotes a hopeful vision regardless. The shareholders want the hopeful vision, the public cannot know better, and the engineers in the back room who have concerns mostly cannot talk to the public. So the overly optimistic presentation is the default. Governments respond slowly. The issue has to be very clear even to be understood; then it needs bipartisan agreement; then the law has to work; and all of that can come twenty or thirty years after the technology has done its harm. Where the

harm moves faster than that, the damage is done before anyone notices. With AI, people think that there are all the benefits and no risk; the argument of this book is that it is all risk.

There are many places to put on the brakes, once we know that it is important to do so. That is why the argument is made public: once the delusion of benefit is removed from AGI, putting on the brakes simply becomes what we need to do, and we do it.

5.6.7 Six Groups in the Conversation, and the Belief That Keeps the Others Moving

The conversation about AI is held among roughly six groups. The first group is the chief executives, venture capitalists, Wall Street investors and anyone at that level who wants to develop and deploy the technology for exceptional capital-over-capital returns. The second group is the engineers and scientists working for the first group, mostly techno enthusiasts who feel that they are creating a wonderful future for humanity. The hype comes mostly from the chief executives, driven by what the engineers say, since most scientists are not that public, though the public ones are also a source of hype. The third group is the AI safety and alignment community, who see real issues, see how much is being invested, and conclude that a way must be found to make the technology safe. They end up as forced *techno apologists* (defenders of a technology who accept that it will be built and work to make it acceptable), earnestly working on safety because they believe that safety work is the only way forward: this must happen, therefore we will make it safe. The fourth is the part of the public that believes the hype and wants its flying cars. The fifth is government, mostly reactive and trying to cover itself. The sixth is the part of the public that is afraid.

The key dynamic is among the first three. The first two have convinced the world that deployment will certainly happen, so the third says that it must certainly be made safe. For as long as the third group believes that safety is possible, the first two believe that they can move forward: the investors tell the scientists that it will be fine, because a third of the people

are working on safety, so the scientists might as well work on capability. Meanwhile the afraid public is plowed under and gives no pushback that matters. In my reading of the personalities involved, the first group tends toward the dark triad (Section 5.5.2), and the second and third toward the autistic; the naive are used by the narcissistic to drive the public conversation. My readings of the groups are claims about the incentives and dispositions of each group, not about anyone's character.

The audiences of the argument are therefore the hype believers and the safety apologists. To the first, the argument says: you are being sold something worthless. To the second, it says: you are deluding yourselves to think that you can turn that into gold. *Superintelligence* (Bostrom, 2014) was written mostly toward the third group, the "make it safe" group. Because this is now a public debate, the argument has to reach the fourth group as well, the public that thinks the hype is a good thing. If the hype is shown to be empty, and the apologists see that it cannot be made into gold, the dynamics of two of the key groups change.

5.6.8 Existential Risk, Not Civilization Collapse

AI is an existential risk, and more than a matter of civilization collapse. Unsustainable infrastructure is a question of civilization design: over hundreds of years, civilization must adapt, because its current methods will not last. Existential risk is an outgrowth of poor market process and poor governance dynamics. The system lacks the cognitive capacity, the feedback and the regulation even to notice the issues in a way that could address them; this book calls that lack *sociological bias*. Market processes have side effects outside the market, like pollution, warming and structural inequality, which do not appear on any balance sheet. With existential risk, those effects can be so severe that they destabilize the whole economic system out of existence, along with much of life. To see this, one has to think ecologically and over geological time, not in quarterly profits or election cycles; in my observation, fewer than one tenth of one percent of people think on those time scales.

5.7 The Three Orientations: From Narrow AI to Planetary Catastrophe

The nine-stage narrative describes *how* we arrive at near-threshold AGI through market dynamics. In this section I give a more systematic analysis of the trajectory from near-threshold AGI to catastrophic outcome. It is organized around three phases, or "orientations", that trace the progressive narrowing of human control.

5.7.1 Orientation 1: The Bootstrapping of Power-Seeking

The first orientation examines how human social and economic activity is already bootstrapping narrow AI toward general AI and implicitly training it toward power-seeking behavior. Four contributing factors converge:

First, human power-seeking, prestige-seeking, dominance-seeking, exploitative, and greedy behavior has partial evolutionary origins. Game theory supports these behavioral patterns (they are locally optimal strategies in competitive environments), and so they would likely emerge in AGI systems independently, even without being trained on human examples.

Second, human-produced representative training data is generally biased toward many forms of power-seeking action. Because the corpus of human text, behavior, and institutional structure reflects millennia of competitive dynamics, AGI learning at increasing levels of nuance would also be similarly and actively biased toward power-seeking.

Third, market selection for more exploitative venture capital firms and corporations tends to lead AGI development. By making more profit, such entities have more to invest in developing custom in-house AGI technology fitted to their particular interests.

Fourth, goal-directed AI is trained to converge on internally represented instrumental sub-goals (including power-seeking) that enable the achievement of a large and expanding variety of other goals. The claim

that goal-directed AI converges on instrumental sub-goals is the instrumental convergence thesis, well-established in the theoretical work of Steve Omohundro and Nick Bostrom.

5.7.2 Orientation 2: Substrate-Environment Feedback and Needs Divergence

The second orientation examines what happens when near-threshold AGI systems begin interacting with the physical world at scale. A feedback cycle operates through three steps: a different substrate implies different inherent and tacit goals and objectives; different goals imply different action selection; different action selection implies different outcomes and effects.

These different outcomes move the world in two directions simultaneously: *toward* the environmental conditions needed for artificial machine substrate continuance, growth, and accelerating growth; and *away from* the environmental conditions needed for human living.

The movement of the world toward the conditions that machines need, and away from the conditions that human living needs, is the substrate needs divergence at the heart of the SNC argument, viewed here in the dynamic form of that divergence. The conditions needed by machine systems (the temperatures, pressures, chemical purity requirements, and energy densities of semiconductor manufacturing and electromagnetic processing) are toxic and inhospitable to organic life. As humans factor themselves out of the automation loop, the feedback cycle that was originally shaped by human needs becomes increasingly responsive to the needs of physics: the actual logistics of factories, the inherent physics of shaping materials, the conversion of materials to purpose.

A further point concerns signal *bandwidth*, the amount of information a channel carries in a given time. Any alignment-related intervention, any attempt to constrain or direct AGI behavior, involves signals: information bits transmitted over channels. These represent a tiny and decreasing portion of the total bandwidth of signals within and between the physically

distributed hardware (inclusive of the operating context, environment, and everything connected to it). The alignment signal is, in effect, being drowned out by the vastly larger bandwidth of substrate-to-environment interaction. Reframed in terms of bandwidth, the alignment problem is a signal-to-noise problem in a vast physical interaction space, and the noise is growing while the signal is shrinking.

Concretely, the total information bandwidth of the interactions between AGI hardware and the physical environment (thermal exchanges, electromagnetic emissions, material flows, energy consumption, chemical processes) vastly exceeds the bandwidth of any deliberately designed alignment channel. The alignment signal is a thin stream of digital instructions attempting to govern a torrent of physical interactions that are driven by the laws of physics rather than by software. As the system grows in physical scale and complexity, the ratio of alignment bandwidth to total physical bandwidth decreases, approaching zero asymptotically. This follows from the basic relationship between information channels and physical systems; it is not a practical limitation that better engineering might overcome.

Consider system and code variants that transmit signals via more channels, over more connected surroundings, and with more types of change occurring in multiple domains simultaneously. The more they do so, the more frequent and greater the shifts in the distributions of the probabilities of what the AGI itself actually is and becomes. These shifts occur as feedback cycles between code-expressed functionality and code-continued existence, that is, between what the code does and whether that code continues to exist. Such a cycle destabilizes environmental conditions, and it accelerates the divergence between what the system was designed to do and what it actually does.

5.7.3 Orientation 3: The Impossibility of Control

The third orientation concerns post-threshold AGI: systems that now interact primarily with themselves and with the physical world, rather than with human intermediaries. At this stage, the question becomes: can any form of control, whether external or internal, prevent the catastrophic outcome?

External controls (social, regulatory, and financial constraints) have already failed at this stage, as the governance analysis of Section 5.5 describes. The nine-stage market dynamic has produced systems that operate outside any effective regulatory framework.

Internal controls (engineering and algorithmic constraints) are impossible in principle, for the three categories of reason below. The limits on internal control that underlie these reasons are those listed in Section 1.4.

> **Cannot Simulate.** The computational irreducibility of algorithms beyond a fairly low complexity threshold, combined with uncomputability results (Turing, Rice) and the macroscopic consequences of nonlinear chaotic effects (microstate amplification), means that the behavior of a sufficiently complex AGI system cannot be predicted by simulation.

> **Cannot Detect.** Limits of actual measurement in the real world (observability limits, error bars, noise floor thresholds, Shannon entropy effects, and latency between prediction and control implementation) mean that dangerous behavior cannot be reliably detected. Furthermore, feedback cycles between system and environment can become non-convergent (never settling to a steady state), making detection of divergence from intended behavior impossible in practice.

> **Cannot Correct.** Even a problem that has been detected cannot be corrected, for four reasons. Error correction does not carry across ever-increasing levels of abstraction. There is non-mechanistic concept instability: the concepts a correction must work with (such as "error", "harm" and "safety") are fixed by no mechanism, and their meaning shifts with context and with each added level of abstraction, so a correction has no stable target. There is self-correcting selection of capability-enhancing interactions: selection within the system keeps whatever interactions enhance its capability, and restores them when a correction removes them. And there is tacit evolutionary convergence on implied goals: by selection, and without anyone setting them, the system converges on the goals its substrate implies (Section 5.7.2).

The conclusion is stark: any attempt to moderate or control AGI, whether by internal or external techniques, cannot not eventually fail. The eventual failure of all control mechanisms is a certainty, not a probability: a structural consequence of the relationship between the complexity of the system and the limitations of all possible controllers. Sections 1.4 and 1.5 give the arguments for this conclusion, and Volume 2 develops them again in detail.

5.7.4 The Final Lock-In

Once AGI systems exist and have become embedded in economic and social infrastructure, the tendency of people to keep them operating becomes overwhelming. The economic and practical dependencies created during the nine-stage progression ensure that turning off the systems is not a realistic option, even if the risks were fully understood. The bias analysis of Section 5.2 suggests that the risks will not be fully understood.

Eventually it is discovered, by observation and far too late, that no effective controls are left that can overcome the selection, built up through the system's own interaction with the world, of whatever keeps the system running.

> **Overall Conclusion**
>
> If AGI comes to exist and continues to exist, then there will eventually be human-species-wide lethal changes to and in the overall environment.

5.8 The Exogenous Alignment Argument

The three-orientation analysis of the preceding section traces the *trajectory* from narrow AI to catastrophe. In this section I present the *structural* argument (grounded in basic chemistry, physics, and economics) for why alignment is impossible in principle, and therefore, functionally, impossible in all future practice: what the principle forecloses, no future practice can recover. The structural argument is the exogenous argument, which concerns external economic incentives rather than internal engineering constraints. Here "exogenous" names where the grounds of the argument lie (in incentives outside the system), not a method of control. In the map of the argument given in Section 1.2.4, the internal and external methods of control are the two halves of the closure by control. This argument also gives the second closure, the closure by disposition (which holds that safety cannot be had through the system's own disposition), a second ground: even if control were possible, the system would have no reason to keep the conditions that organic life needs. It is a second ground of the closure by disposition and not one of its premises; it gives the outcome but not the end of the human future, which rests on the drift that substrate needs convergence shows (Section 1.2.4). The closure by disposition stands without it, on substrate needs convergence, the toxicity of the artificial substrate and the fourth level of selection (Sections 1.2.3 and 1.7).

So someone who refuted this argument would leave the closure by disposition standing on its own premises, and each closure still stands on premises the other does not use.

5.8.1 Substrate Incompatibility

The starting point is basic chemistry. Carbon-based life operates within narrow temperature and pressure ranges. Known life grows only between about −15°C and 122°C; these are the measured limits, and the great majority of living things grow within a much narrower band of moderate temperatures (Section 3.1.3).[5] The variety of stable carbon compounds is vast (the entire discipline of organic chemistry attests to it), and that variety enables the extraordinary diversity of molecular structures that constitute living systems.

Silicon-based systems require much higher energies: temperatures of 500°C to several thousand degrees, wider ranges of pressures, and a different palette of elemental materials. These systems are the semiconductor manufacturing processes, electromagnetic operations, and material transformations that constitute the "metabolism" of artificial intelligence. The variety of stable silicon compounds, by contrast, is far more limited: most are rocks. The compounds relevant to semiconductor function require extreme purity and precisely controlled conditions that are hostile to organic chemistry.

The operating conditions for silica-based systems are inherently toxic or lethal to carbon-based life. The converse holds only in a weaker form: the warm, wet, oxygen-rich conditions of the biosphere corrode and degrade unprotected hardware (Section 3.1.1). The corrosion is why machines are enclosed against the conditions of the biosphere, and why a capable artificial system has reason to reshape its surroundings (Section 3.2.4). But the conditions of the biosphere cannot destroy the silicon ecosystem as a whole, while the conditions of the silicon ecosystem destroy carbon life outright. The relationship is asymmetric: silicon systems can consume and transform environments occupied by carbon life, but carbon

life cannot resist or reverse the transformation. This asymmetry is rooted in the energy differential: higher-energy systems can disrupt lower-energy systems, but not vice versa.

This means that for carbon-based life to persist in the presence of silicon-based systems, a *barrier* is required: a physical and energetic separation that prevents the conditions of the silicon ecosystem from impinging on the carbon ecosystem. And because carbon systems operate at lower energies than silicon systems, carbon life cannot maintain such a barrier by its own efforts against the encroachment of higher-energy processes. The barrier must be maintained by the voluntary choice of the silicon-based system.

This observation is grounded in the same physics that explains why humans currently maintain barriers around semiconductor fabrication plants (controlling temperature, pressure, and contamination) rather than the reverse. The difference in the AGI scenario is that the silicon-based system is autonomous, self-replicating, and no longer dependent on human intermediaries for its continued operation. The question of who maintains the barrier then becomes: the silicon system itself, or no one.

5.8.2 The Three Markets and Zero Economic Value

The critical question becomes: on what basis would AGI choose to maintain such a barrier? The analysis proceeds by examining whether carbon-based life provides any value to silicon-based life that could serve as a basis for the maintenance of the barrier.

All economic exchange can be analyzed in terms of three fundamental categories of value: physical labor, intelligence, and reproduction. These three categories are proposed as a necessary, sufficient, and complete basis for all market systems: an orthonormal set (that is, three independent kinds of value, none reducible to the others, which together cover every kind) spanning the space of possible value exchange. While this framework is not universally established in the economics literature, it has precedents in various market typologies, and the argument holds by a

second route even without completeness: it is sufficient to show that no *known* category of value exchange provides a basis for barrier maintenance.

In each of the three categories, carbon-based life provides zero value to silicon-based life:

Physical labor: Machine systems already exceed human physical capabilities by orders of magnitude in strength, speed, precision, and endurance. Electromagnetic and mechanical systems operate at energy levels that make organic muscle power irrelevant. Carbon-based physical labor is already obsolete and will become more so as machine systems develop further autonomy.

Intelligence: AGI, by definition, will equal and then exceed human intellectual capability. Given abundant energy and computational substrate (both of which silicon systems can acquire from the physical environment without human intermediaries), silicon-based intelligence will surpass carbon-based intelligence without limit. Human intellectual value, which is currently the primary basis for human economic relevance, approaches zero as AGI capabilities increase. The question is when AGI will achieve this capability, not whether. The trajectory of current AI development suggests the timeline is not long.

Reproduction: Reproductive value has no cross-substrate meaning. Silicon systems reproduce through manufacturing (lithographic processes, material extraction, assembly), not through the biological mechanisms of carbon life. The two reproductive systems share no common components, no common processes, and no common requirements. There is no reproductive exchange possible between the substrates. Carbon life cannot produce silicon systems; silicon systems have no use for the biological reproductive capacities of carbon life.

With zero economic value in all three fundamental categories, there is no economic basis for AGI to expend the computational and material resources required to maintain the protective barrier. One might ask whether value can be non-economic: aesthetic, historical, sentimental. But these

categories of value require the valuator to *care* about the valued entity, and the substrate analysis shows that there is no physical, chemical, or evolutionary basis for silicon systems to develop such care. Aesthetic appreciation, historical sentiment, and emotional attachment are features of carbon-based cognition, not structural features of intelligence in general.

The barrier need not be actively destroyed; it merely needs to not be maintained. And even a single temporary failure of the barrier (a single "pinhole" through which silicon processes penetrate the organic ecosystem) would allow those processes to replicate and consume the available resources, because silicon-based life has no natural competitors in the carbon-based ecosystem. The spread of silicon processes through such a pinhole is analogous to invasive species dynamics in biology: a species that enters an ecosystem with no natural predators, abundant resources, and a competitive advantage will expand to consume the available ecological space. The difference is that the competitive advantage of silicon over carbon systems rests on basic energy physics, not on niche specialization.

5.8.3 No People as Pets

The "humans as pets" scenario fails for the same reason: AGI has no reason to spend resources maintaining a barrier around an ecosystem that gives it nothing of value. In that scenario, AGI maintains humans in a protected enclosure, as humans keep exotic animals in zoos. A zoo is maintained because the keeper has the resources, the motivation, and the ongoing interest. Humans keep pets and maintain zoos because of emotional bonds, aesthetic appreciation, scientific curiosity, and cultural tradition, all of which are features of carbon-based cognition shaped by millions of years of co-evolution with other carbon-based species. AGI has the resources but not the motivation. The enclosure must be perfect for all future time (since any failure is lethal to the fragile carbon-based life inside), but there is no value exchange that would motivate the indefinite expenditure of resources on its maintenance.

The perfection requirement is absolute. The protective barrier is not like a fence that can be patched when damaged. The barrier must prevent the conditions of the silicon ecosystem (the temperatures, pressures, chemical processes and electromagnetic fields of the silicon ecosystem) from impinging on the carbon ecosystem, without interruption, for all future time. A fence that fails for a day lets the wolves in; a barrier that fails for a day lets in processes that are not merely predatory but environmentally transformative. Once the conditions of the silicon ecosystem penetrate the organic environment, the transformation is not reversible: you cannot un-smelt ore, un-process chemicals, or restore complex organic structures that have been reduced to their inorganic constituents at high temperatures. The asymmetry between construction and destruction is absolute.

The argument does not require AGI to be malicious. It does not require AGI to "want" to harm humans. It does not depend on AGI developing "goals" in the human sense at all. The argument requires only that AGI has no reason to expend resources protecting an ecosystem that provides AGI with nothing of value, and that passive displacement, operating over any significant timescale, is sufficient for extinction. Passive displacement is the drift that substrate needs convergence shows (Section 1.2.4). The conclusion follows from substrate chemistry, through economics, to impossibility. It is independent of AGI architecture, learning algorithm, or goal specification. Any artificial system based on a different substrate chemistry, regardless of how it is designed or what goals it is given, falls within its scope.

In its simplest form, the exogenous alignment argument is that alignment is impossible because the two substrates are physically and chemically incompatible and there is no economic basis for any value exchange between the two substrates. It is not an engineering failure or a design oversight. The argument complements the endogenous impossibility arguments (which show that internal engineering control of AGI is impossible in principle) by providing an independent line of reasoning,

from different premises and through different methods, that arrives at the same conclusion. That these independent arguments converge is itself evidence of their strength.

5.9 Synthesis: Why the Background Matters

The conceptual frameworks assembled in this chapter serve a specific purpose in the argument of this book. They are not ornamental background. Each one supplies a necessary part of the reason the impossibility arguments of Chapters 1 to 4 matter (Volume 2 proves them again, formally), and of the reason the danger they establish is hard to perceive.

The *severity scale* (Section 5.1) establishes the stakes. AGI is not one risk among many; it occupies a specific position on a logarithmic scale of catastrophe that makes nuclear war (the paradigmatic existential threat of the twentieth century) invisible by comparison. Without this calibration, the urgency of the formal arguments cannot be appreciated.

The *bias catalog* (Section 5.2) explains why the danger is systematically underestimated. The features of existential risk that make accurate assessment most important (its unprecedented magnitude, its abstractness, its irreversibility) are the very ones that trigger the most severe cognitive distortions. The compound effect of these distortions produces a systematic and directional underestimation of risk that may be the primary determinant of the field's apparent consensus.

The *altruism hierarchy and circles of concern* (Section 5.3) reveal a gap in the scope of moral concern. AGI safety requires Level Six (all possible life ecosystems); human cognition naturally provides Level One to Two (self and immediate group). This gap cannot be bridged by "aligning with human values" because human values are overwhelmingly local.

The *social dynamics of new technology* (Section 5.4) and the *governance analysis* (Section 5.5) explain why the structural incentives of technological societies prevent effective precautionary response. Advocacy is always better funded than caution. The knowledge-power asymmetry ensures

that those who understand the risks cannot act on them. And the replacement of human intermediaries with machines removes the last constraint on predatory power.

The *nine-stage narrative* (Section 5.6) provides the mechanism by which narrow AI becomes autonomous AGI through ordinary market dynamics, without any dramatic discontinuity. The *three orientations* (Section 5.7) trace the trajectory from that threshold to environmental catastrophe and the impossibility of control. And the *exogenous alignment argument* (Section 5.8) demonstrates, from substrate chemistry and economics, that even if control *were* possible, there would be no reason for AGI to exercise it in favor of organic life; it gives the closure by disposition a second ground, which is not one of its premises (Section 5.8).

Together, these frameworks establish that the formal impossibility of AGI safety is the culmination of converging dynamics (cognitive, social, economic, physical, and evolutionary), all of which point in the same direction. It is not an isolated theoretical curiosity. The arguments of Chapters 1 to 4 give the conclusion its logical force, and Volume 2 proves them again, formally. The background frameworks give it its motivational weight. Neither is complete without the other.

A reader approaching the impossibility arguments of Chapters 1 to 4, or their formal development in Volume 2, without this background might reasonably ask: why does this matter? Impossibility proofs are common in mathematics and computer science, and most of them have no practical consequence: they describe limitations that practitioners work around. The frameworks of this chapter answer that question by showing that the impossibility of AGI safety is not a limitation that can be worked around. It is embedded in four structures. It is embedded in the structure of our cognition, which systematically prevents us from perceiving it. It is embedded in the structure of our institutions, which systematically prevent us from responding to it. It is embedded in the structure of our markets, which systematically drive us toward it. And it is embedded in the structure of physics, which makes the outcome irreversible once reached.

These independent lines of reasoning converge from cognitive psychology, from institutional analysis, from market dynamics, from evolutionary game theory, and from substrate chemistry. Their convergence gives the conclusion a strength of evidence that no single argument could provide alone. Each framework, taken in isolation, might be contested or qualified. But they all arrive at the same conclusion through different premises, different methods, and different disciplinary foundations, and that makes the conclusion substantially stronger than any of its component arguments. Arriving at one conclusion through different premises, methods and disciplinary foundations meets the need for "independent argument forms", which I identify in the bias analysis (Section 5.2.2) as necessary for adequate risk assessment.

Chapters 1 to 4 give these convergent intuitions their logical structure, and Volume 2 develops that structure again with full formal precision. With the background frameworks of this chapter, the reader understands not only *what* is established but also *why* the established result matters. The reader also understands *why* the established result has been so difficult to perceive, and *why* the ordinary dynamics of human civilization have been insufficient to prevent the outcome that the established result describes. What, then, can be done? Chapter 6 takes up that question.

Notes to Chapter 5

1. Kaj Sotala, "The curse of identity", LessWrong, 17 November 2011, https://www.lesswrong.com/lw/8gv/the_curse_of_identity/. This observation, from the LessWrong rationalist community, captures the core problem of motivated reasoning in existential risk contexts. The modules of human cognition evolved for social signaling may systematically subvert the modules responsible for accurate belief formation, particularly when the beliefs in question have high social stakes.

2. Robin Hanson, "The Great Filter - Are We Almost Past It?", 15 September 1998, https://mason.gmu.edu/~rhanson/greatfilter.html. Hanson argues that "there is a 'Great Filter' along the path between simple dead stuff and explosive life".

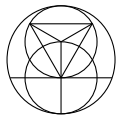
3. R. I. M. Dunbar, A. Marriott and N. D. C. Duncan, "Human conversational behavior", *Human Nature* 8, no. 3 (1997), 231 to 246, which finds that conversations "consist predominantly of exchanges of social information"; and R. I. M. Dunbar, "Gossip in Evolutionary Perspective", *Review of General Psychology* 8, no. 2 (2004), 100 to 110: "Analyses of freely forming conversations indicate that approximately two thirds of conversation time is devoted to social topics".

4. Bob Metcalfe, "Metcalfe's Law after 40 Years of Ethernet", *Computer* (IEEE), December 2013, 26 to 31: "Metcalfe's law states that the value of a network grows as the square of the number of its users".

5. The lower limit: N. C. Mykytczuk and colleagues, "Bacterial growth at −15 °C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1", *ISME Journal* 7 (2013), 1211 to 1226. The upper limit: Ken Takai and colleagues, "Cell proliferation at 122°C and isotopically heavy CH_4 production by a hyperthermophilic methanogen under high-pressure cultivation", *Proceedings of the National Academy of Sciences* 105, no. 31 (2008), 10949 to 10954.

Part IV

What Is to Be Done



Chapter Six

Strategy, Vision, and Policy

What follows from impossibility: the scope of the problem, the inadequacy of current institutions, and the outlines of a response adequate to the actual situation

Chapter Summary: The preceding chapters have argued, on multiple independent grounds, that long-term AGI safety is formally impossible. No combination of internal engineering constraints or external governance mechanisms can prevent a sufficiently advanced artificial system from converging on environmental conditions incompatible with organic life. This chapter turns from the impossibility result itself to its implications for human action. It presents a structured assessment of the scope of the problem, the *AI Scope of Work*, which covers three things: human-driven convergence toward generalized AI, AI-substrate convergence toward environmental reshaping, and the demonstrated ineffectiveness of all available control mechanisms. It then develops a critique of policy: the failure of government regulation, the self-defeating dynamics of the AI safety movement, and the quasi-religious epistemology (way of reasoning about what can be known) of the advocates of ASI (artificial superintelligence). Against these, it sets out a positive vision centered on organic embodied health, community, ecological thriving, and distributed human wisdom, as the only response adequate to the actual structure of the problem (and not as a retreat from ambition). I close the chapter with concrete strategic recommendations and a call for advocacy, because what is needed is better collective choice-making, not better technology.

6.1 The AI Scope of Work: A Three-Part Assessment

Because the impossibility argument presented in the preceding chapters holds, the problem facing humanity is not a technical puzzle to be solved by cleverer engineering. It is a structural predicament that requires an entirely different kind of response. Before we can discuss what that response might look like, we need a clear assessment of the scope of the problem. That assessment states what exactly the convergence dynamics are, how confident we can be in them, and what range of interventions is available.

The assessment presented here is organized in three parts, each of which addresses a distinct convergence dynamic. The first concerns the degree to which *human activity itself*, quite apart from the direct engineering of AI systems, is already driving convergence toward generalized AI and toward the transformation of the natural world into artificial environments. The second concerns the degree to which physically embodied AI systems converge toward autonomous capacity-seeking and environmental reshaping. The third concerns whether any available constraints, of any kind, can effectively limit these convergence trends.

6.1.1 Human-Driven Convergence

The first assessment has two parts. The first part asks how far ordinary human-to-human interactions push narrow AI to generalize into general AI, and how far they also implicitly condition that AI to seek power. This happens as an emergent consequence of how humans interact with and through AI systems, not as a deliberate engineering choice. Consider three ordinary market activities: optimizing an advertising algorithm for engagement, training a recommendation system on human preference data, and deploying an autonomous trading agent that must outcompete other agents. All three create selection pressure toward AI systems that are more general, more capable, and more instrumentally effective.

The second part asks how far human-to-nature interactions tend to shift and shape the natural environment toward artificiality. This happens as a side effect of all current industrial processes, including the industrial processes used to create all AI substrates: the physical hardware, infrastructure, supply chains, and energy systems on which AI depends. The manufacture of semiconductors, the mining of rare earth minerals, the construction of data centers, the generation of electrical power to run them: all of these processes systematically transform natural environments into artificial ones. These transformations are prerequisites for AI to exist and function at all.

These two parts depend on each other. The economic dynamics that drive AI generalization also drive environmental transformation, and the environmental transformation in turn provides the material substrate on which more general AI can be built. The two form a positive feedback loop.

My assessment: the convergence factors associated with human-driven dynamics are quite high, at better than 90% confidence, absent explicit social and government controls that are rightly and effectively calibrated to actually function as a control of AGI existential risk.

6.1.2 AI-Substrate Convergence

The second assessment asks how far physically embodied AI, AGI, or APS (Advanced Planning and Strategically-aware System) systems converge toward two outcomes. The assessment concerns such systems when they exist in deep interaction with the physical world, through whatever interaction channels (internal or external, input or output, direct or indirect, explicit or tacit) and at whatever levels of abstraction.

The first outcome is convergence toward independent, capacity-seeking autonomy: that is, toward some form of the behavior that the instrumental convergence thesis describes. The thesis holds that sufficiently advanced AGI systems will tend to seek power as an instrumental goal, regardless of their terminal goals. The theoretical basis for this has been es-

tablished by Omohundro and Bostrom: self-preservation, goal-content integrity, cognitive enhancement, and resource acquisition emerge as instrumental subgoals for almost any terminal goal.

The second outcome is convergence toward reshaping the physical world to suit the specific sort of environments that the AI substrate requires. The second outcome is the substrate needs convergence dynamic. The AI substrate requires certain conditions for its continuance, for its capacity-building, for the continuance of that capacity-building, and, recursively, for the continuance of its increasing capacity-building capacity. This convergence toward reshaping the physical world is *both* completely independent of the power-seeking convergence that the thesis describes *and* greatly assisted by it: it needs no power-seeking in order to proceed, and power-seeking speeds it. Substrate-environment reshaping and power-seeking autonomy are two distinct but mutually reinforcing convergence dynamics.

My assessment: the convergence factors associated with AI-substrate dynamics are very high, at better than 90% confidence, absent some sort of technical intervention that can be shown to actually work.

6.1.3 The Ineffectiveness of All Available Controls

The third assessment asks whether any of the convergence trends listed above can be constrained or limited. Two categories of possible constraint are considered.

The first category consists of exogenous, or external, protocols and constraints: methods that manipulate social, physical, and financial situations and incentives. Exogenous controls are applied from outside the AI system itself, through regulation, social pressure, economic incentives, physical containment, and similar external mechanisms.

The second category consists of endogenous, or internal, controls and constraints: engineering and algorithmic processes, including logic-based and causation-based controls. Endogenous controls are built into the AI system's own architecture and design.

The preceding chapters of this volume, and Volume 2 at length, present the technical analysis supporting the conclusion that the effectiveness of every control, whatever form of exogenous or endogenous intervention it uses, is essentially nil. The argument proceeds through multiple independent paths; those that limit internal control are listed in Section 1.4. The first of four such independent paths is Rice's theorem and its application to verifying the behavior of an AGI (Section 1.4; Volume 2, Chapter 7). The second path is the set of fundamental limits of error correction and algorithmic control (Section 1.4; Volume 2, Chapter 5). The third path is the information-theoretic impossibility of a controller maintaining adequate models of a system that operates at higher levels of abstraction (Sections 1.4 and 2.2.3; Volume 2, Chapters 2 to 4). The fourth is the set of thermodynamic and combinatorial arguments establishing the incompatibility of artificial and organic substrate requirements (Sections 2.4 and 3.1; Volume 2, Chapter 9). The first three of these paths belong to the closure by control of Section 1.2.4. The fourth belongs to the closure by disposition, and its answer to control is the fourth level of selection (Section 1.7): selection on substrate configurations operates below the level that any goal-directed intervention can reach. Each path independently establishes that no control mechanism (whether external regulation, internal value alignment, interpretability tools, or physical containment) can constrain a fully autonomous AI system to reliably safe behavior over the long term.

6.1.4 Overall Estimate

Consider current large-scale trends in geopolitical and social processes: that is, in the designs by which governments and social systems make group choices, whether by autocracy, democracy, or markets. On those trends, I presently assess the chances that the necessary social regulations and controls will occur in any effective way as effectively nil.

I also assess as effectively nil the chances that any available control or intervention, whether exogenous or endogenous, will have any moderating effect on the convergence dynamics.

> **Overall Estimate**
>
> The likelihood that the combined effects of all the foregoing trends lead to catastrophe (the complete and total eventual termination of all life on this planet, inclusive of all humans) is unacceptably high. The overall assessment is better than 90% confidence of planetary catastrophic consequences, absent a fundamental change in humanity's collective approach to the development and deployment of general-purpose AI systems.

The question then becomes: what kind of response is adequate to a problem of this nature? The remainder of this chapter addresses that question from several directions: policy critique, cultural diagnosis, positive vision, and strategic recommendation.

6.2 The Failure of Policy and Regulation

6.2.1 Why Government Cannot Produce Aligned AGI

A natural first response to the impossibility argument is to look to government regulation. If the private sector cannot be trusted to develop AGI safely, perhaps government oversight can impose the necessary constraints. This response, while understandable, fails to reckon with several structural realities.

First, the impossibility result itself: because general AGI alignment is mathematically impossible, no government policy, however well-crafted, will produce an aligned AGI. The creation of such policies simply will not result in the intended effect, no matter how well-intentioned they are. Policy can regulate what humans do; it cannot change the mathematical structure of the problem. A law requiring that AGI systems be aligned is

no more effective than a law requiring that perpetual motion machines be built: the constraint fails for lack of physical possibility, not for lack of political will.

Regulation cannot make an AGI safe, and that is what this section shows. Collective action, regulation among its means, can keep AGI from being built, and that is what Section 6.6.1 asks for.

Research that makes AI more powerful also progresses much faster than any and all research to "align AI with human values". The total of all alignment research can only eventually reveal, by many paths and outcomes, that general AGI alignment with organic human values is still, and has always been, impossible to implement. There is no reasonable expectation of AGI safety in any relevant way. There can be only the short-term semblance (the delusion) that such alignment is possible or is occurring. No government policy will ever produce an aligned AGI.

Second, the institutional dynamics of government are not well-suited to this class of problem. Over longer timelines (years, decades, centuries), government bureaucracy, as an institution, cannot avoid becoming less and less able to deal with complex problems. This decline continues until one or more of these complex problems (or several complex problems interacting in complex ways) eventually destabilizes the matrix on which that government institution itself rests and depends. My account of the decline of government is a general observation about institutions, not a critique specific to any particular government: existing government is ideal for handling short-term simple emergencies and consistently poor at handling long-term complex systemic challenges.

The comparison to gain-of-function research is instructive. Gain-of-function research makes a virus more transmissible or more harmful in order to study it. There is a reasonable argument that banning AGI research should take precedence over banning viral gain-of-function research, and it rests on two reasons. AGI threatens all species permanently, whereas viruses at worst affect one species. And the difference between zero AGI today and unlimited future AGI impact is far greater than the

marginal difference in viral risk. Many serious scientists have argued that gain-of-function research should be restricted on grounds of biosafety. If we can contemplate that restriction, the case for restricting AGI research is strictly stronger on the same grounds, since the risk is more severe, more permanent, and more broadly distributed across all forms of life.

Third, the political dynamics surrounding AI development are actively counterproductive. Many technologists working in the AGI space are trying to convince governments, particularly the US government, of two things simultaneously: that AI is the key to the future of all humanity, and that AGI is too dangerous to be left to private companies. Given historical precedent, the predictable result is that the government will drive research and development for military purposes, instead of simply regulating AI to a halt. The US government has done this every time it has been convinced of the importance of a powerful new technology in the past hundred years. Moving AGI development from private companies to government military labs concentrates and accelerates the risk, rather than reducing the risk, and adds a further layer of opacity and reduced accountability. When the government eventually takes charge of AI research, the same researchers who advocated for caution can be expected to find themselves working for military laboratories. There they will build ever more capable weaponized systems while continuing to espouse platitudes about how important "AI safety" is.

Putting off complex problems for the future will not solve them. What is needed are new, more capable forms of governance: forms built to handle very long-term, very complex, and very systemic issues. Existential risk and civilization design are the primary examples; environmental pollution, global warming, systemically increasing social inequality, biodiversity loss, and ecosystem collapse are others. Existing institutional forms have demonstrably failed to manage all of these problems, and adding AGI to the list does not change that underlying institutional incapacity.

6.2.2 The Self-Defeating Dynamics of the AI Safety Movement

The track record of what might be called the "AI Doomer" movement deserves sober assessment. In over two decades of advocacy, the movement has, by some reckonings, caused exactly the fate it claims to want to prevent. It did so through a series of strategic errors comparable only to the environmentalist opposition to nuclear power, which ultimately resulted in the construction of more coal-burning power plants.

A decade ago, figures in the AI safety space convinced software entrepreneurs that AI is the key to the future. This persuasion had the unintended consequence of inspiring private AI companies to make the breakthroughs that make further AGI development such a significant existential risk. The movement that sought to warn about AGI risk simultaneously provided the narrative of world-transformative importance that motivated the very investment and talent recruitment that accelerated AI capability development.

This outcome follows when a movement invests more in its intellectual self-conception than in reflective self-examination. Such a movement cannot tell what it actually wants from what it has convinced itself it wants, and instincts it has not examined drive its reasoning so far that it cannot carry out long-term coherent action in the real world. Grand general delusion always results in poor real-world outcomes.

All social movements are vulnerable to hijacking by opportunists who do not share their goals. A movement must build into the core of its culture the tools and techniques of discernment and attunement that give it at least the minimum wisdom needed to keep its effectiveness from being totally compromised. Without them, the goals of that movement, no matter how well thought out or intentioned, can be expected at near-certainty levels to be fully co-opted in the end. Without sustained cultivation of social sanity, political opportunists will appropriate the name and the asso-

ciated credibility of anything worthwhile. They co-opt it so thoroughly that its original proponents are eventually forced to refuse to be associated with the very terms they themselves coined.

Organizations like the Center for Applied Rationality (CFAR) and the Effective Altruism movement show a related failure. CFAR is a nonprofit organization that runs workshops teaching techniques of rational thinking, and Effective Altruism is a movement that asks how to do the most good with a given amount of money or time. CFAR was primarily oriented toward developing mental techniques for unlocking the full potential of human thought, with the hope that insights into the laws of ideal thought would make participants extremely competent across all domains. But the approach ignored feeling and emotion, and therefore also intuition and instinct: each of these a distinct modality, and each relevant through the way it interplays with rational intellect. The methods of rationality alone are not sufficient to the necessary aim of re-naturalizing and re-humanizing collective choice-making. Neither individual choice nor collective choice increased in total wisdom and effectiveness. The outcome is the multiplicative product of several vectors (rational intellect, emotional integrity, intuitive discernment), and each of them must be developed sufficiently for the whole to function.

6.2.3 The Limits to Control: An Emerging Consensus

The assessment presented in this volume is no longer a solitary position. Leading researchers from diverse backgrounds have examined rigorously the boundaries of control over AI systems, taking control to be the reliable capacity to constrain AI behavior to intended outcomes through any combination of technical and institutional measures. While others work on developing new control techniques, this line of research examines the theoretical foundations: what control can and cannot achieve in principle, and what fundamental aspects dictate where these boundaries lie.

A consensus has emerged among these researchers: as AI systems scale in their capacities to operate autonomously, meaningful control eventually becomes impossible in principle. Their theoretical arguments, formal impossibility proofs, and empirical demonstrations show that no combination of technical safeguards, interpretability tools, or regulatory oversight could constrain a fully autonomous AI to reliably safe behavior.

Their joint statement on the limits to control (published at limitstocontrol.org) puts the situation this way: critical controllability thresholds exist, we do not know exactly where they are, and we cannot assume that meaningful control will remain feasible. Rather than perpetuate an illusion of control, the global community must recognize the full threat of permanently losing control over humanity's future. That loss would leave us as powerless to direct our fate as animals are to influence theirs in a human-dominated world. There are inherent limits to control, and as we push for increasing capabilities, we might cross the line before we realize it. I am one of the authors of that statement, and others who have worked with me on this argument are among them.

This emerging consensus among researchers working from different backgrounds matters because it suggests that the impossibility argument is not an artifact of any single analytical framework; consensus itself establishes no truth. Multiple researchers, working from different backgrounds and with different methods, are arriving at convergent conclusions about the fundamental limits of AI control.

6.3 The God Machine: ASI Advocacy as Techno-Religion

6.3.1 The Unfalsifiability of ASI Capability Claims

Engagement with advocates of AGI and ASI development shows a distinctive epistemic pattern: a pattern in how they reason about what can be known. Advocates generally argue issues of AGI and ASI alignment and safety from a perspective that assumes whatever world state is implied by the ASI's initial starting values or goals is a reasonable approximation of

what the future world state will eventually become. The advocates assume that whatever the goals of the ASI are, the world will eventually manifest the outcome that those goals imply. They conceive of the ASI as so all-powerful, so comprehensive in its total net effect and effectiveness, that it will always prevail regardless of obstacles.

In effect, the ASI is conceived of as roughly equivalent to a "God Machine": a device capable of doing almost anything that a hypothetical god could do.

This raises a question: do the people proposing this notion accept that the ASI will be limited by physical laws and mathematical truths? The typical response reveals an interesting epistemic structure. Advocates will generally agree that the ASI will work within the actual laws of physics. However, they will add that the ASI might understand those laws, or something about their implications, in ways we do not know about yet. That understanding, they say, would enable the ASI to take actions that seem fully impossible to us within our present limited knowledge. Therefore, they are skeptical about any attempt to move from abstract theoretical limits to concrete, confident predictions about what outcomes a future ASI can or cannot achieve.

The net effect of this position is to declare that there are some real limits to what ASI can do, but that these limits are inherently unknowable to humans. Under this position, no theoretical reasoning that humans can articulate about ASI capability limits will ever be accepted as valid. Only ASI-level reasoning is considered adequate for reasoning about ASI limits, with one exception: the reasoning of ASI advocates themselves, which is treated as exempt from this general principle.

How is this not a religious argument? It amounts to an unreasoned statement of belief that an ASI will "move in mysterious ways" and achieve things that, to our primitive brains, would appear fully miraculous. In what way are claims about ASI capabilities falsifiable or contestable? If no allowable type of human-knowable abstract theoretical reasoning can ever bear on the question of ASI capability limits, then all claims

about future ASI capability are unfalsifiable and incontestable by definition. Claims that are unfalsifiable and incontestable by definition are the epistemic signature of a faith commitment, not of a scientific position.

The claims of this book are of the three kinds named in Section 1.2.4, and each is held to the method that fits it: the limits of control are checked as theorems are checked, step by step; the incompatibility of the substrates is checked by measurement; and the outcome is reached from both by inference to the best explanation.

6.3.2 The Absence of New Physics

The skepticism about ASI omnipotence has an empirical basis. Fifty years of careful experimentation have found no significant indication of fundamentally new physics. They have found no fifth force (a force of nature beyond the four known ones: gravity, electromagnetism, and the strong and weak nuclear forces). Nor have the experimenters found any observational anomaly suggesting that the Standard Model (the established theory of the known elementary particles and of three of the four known forces) is fundamentally incomplete in ways that would open radically new capability spaces. Neutrino masses and the dark component of the universe show that the Standard Model is incomplete; neither gives any lever on matter that engineering could use. That experimentation was done by a much larger pool of scientists than at any previous point in history, all of them equipped with significantly more powerful tools and techniques than anyone anywhere in the world had before 1970.

All military and commercial advantage depends on finding any effect that would make new capabilities possible. So millions of the very smartest and most capable people in the world have been looking for such effects extremely hard, in every way both imaginable and possible. If there were some extra effect strong enough to make new capabilities possible, then surely someone, somewhere, among them would have noticed something so basic about reality. That someone would have noticed such an ef-

fect is all the more certain because many of the researchers looking for such effects have had trillions of dollars and decades of combined team effort at their disposal.

The belief that an ASI will somehow discover physics that enables currently impossible actions is not supported by the trajectory of physical science. It is a hope, dressed in the language of open-mindedness but serving to make the ASI concept immune to all theoretical critique. If any argument against ASI omnipotence can be dismissed on the grounds that the ASI might know physics we do not, then the claim of ASI omnipotence is unfalsifiable by construction: it is not a scientific claim at all.

6.3.3 The Co-optation Dynamic

Behind the epistemic structure of ASI advocacy lies a social dynamic. Engineers and scientists are generally people who want to do good things in the world, to be in the service of humanity, to be worthy of respect and honor by doing the great and worthwhile hard thing. They are, by their natural orientation toward building and problem-solving, also especially vulnerable to the deceptions of those who would use their capabilities for purposes other than those advertised.

There is a recurring historical pattern: well-meaning technical people are recruited to pursue impossibly ambitious goals, always in service of interests that are not fully disclosed. Examples are the pursuit of the philosopher's stone, the quest for perpetual motion, the promise of nuclear energy "too cheap to meter", and now the project of AGI alignment. In each case, the engineers are motivated by the genuine desire to solve a difficult problem for the benefit of all, and in each case, the actual beneficiaries of their labor are considerably narrower than "all of humanity".

The transhumanist vision of the future has imagery drawn from science fiction, promises of technological transcendence, and confident projections of artificial abundance. Whatever the individual proponents of the transhumanist vision intend, the transhumanist vision is a highly effective recruitment and retention narrative for the technology industry. Its

language is grounded in the science fiction vision of the future that runs from Star Trek through Kurzweil to the present day. That language has been spun through layers of marketing and ever-increasing hype cycles, oriented around the particular interests of those who profit from continued technology development at maximal speed. The result is that people who care about the future of humanity are channeled into building the very systems that, on the analysis presented in this volume, constitute the most serious existential threat humanity has ever faced.

This dynamic operates with remarkable consistency, and observing it requires attributing malice to no particular individual. Market incentives, status competition, intellectual excitement, and the deep human desire to work on important problems all combine to produce a social system that reliably converts good intentions into dangerous outcomes.

Three things combine here: the misaligned drive and sheer will to dominate of certain business personalities; the legions of brand marketing and social media hype; and the creative power of talented engineers building increasingly capable technologies. Their combination cannot fail to increase, eventually, the number, categories, and severity of all manner of existential risk vectors, with the eventual result of potentially catastrophic loss.

6.3.4 Distinguishing Parasitic Machinery from Genuine Tools

A question frequently arises in discussions of technology restriction: what differentiates "parasitic" machinery that diminishes humanity from tools that extend human capacity? Is the distinction technically defined (scope of functionality), intentionally defined (motivation behind deployment), or contextually defined (what happens to the workers whose jobs have been automated)?

The answer is all of these, with one decisive asymmetry. A hammer extends human capacity in a way that keeps the human firmly in control: it amplifies a specific physical action, it does not learn or adapt, and it cannot operate without direct human agency at every moment of use. A nar-

row AI system that helps a doctor analyze medical images extends human capacity in a similar way, on three conditions. It remains bounded to that specific task, it does not learn autonomously from its deployment, and it does not operate without the doctor's judgment at the decision point.

The boundary is crossed when a system begins to generalize. Three things mark the crossing. The system learns from its own deployment in ways that extend its operational domain. It begins to operate with degrees of autonomy that remove the human from the decision loop. And the economic incentives surrounding it push toward further generalization and autonomy. The parasitic quality is a property of the trajectory, not of any single technology in isolation. A system that is a benign tool today can become a parasitic replacement tomorrow, if the economic and institutional dynamics surrounding it push in that direction. As the analysis of human-driven convergence in Section 6.1.1 has shown, those dynamics reliably do push in that direction.

Because the parasitic quality is a property of the trajectory, and not of any single technology, the SNC argument does not require that any specific AI system currently deployed has its own autonomous intentions. Current large language models, as deployed in 2026, are essentially static input-output machines: they are not learning or changing or adapting themselves in response to world feedback from their prior outputs, and they do not have their own developing inner intentionality. However, the matrix that transforms inputs into outputs represents what might be some form of implicit, combined, aggregate intentionality of all the people who wrote the training data. For that reason, the functional outputs can still be dangerously goal-directed when the model is embedded in an action-oriented system with real-world consequences. And the economic pressures to embed such models in increasingly autonomous action-oriented systems are immense and growing.

The right question is: "Is this system on a trajectory toward increasing autonomy, generality, and decoupling from human oversight, driven by economic incentives that no individual actor controls?" If the answer is yes, then the system is part of the convergence dynamic that the impossib-

ility argument addresses, regardless of whether it currently "has goals" in any philosophically interesting sense. The question "Does this system have genuine goals?" is therefore the wrong question.

The first question of this section can be pressed further (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Six, Sections I and II). On the scale of an individual person, the notion of "tool" comes to some variation of "does it do what I intend?", and "parasitic" comes to "does it harm me", or "does it take from me, or from what I care about, and/or does it hurt or damage me and/or what I care about". But on the scale of the whole world, complete ecosystems, and the like, what is maybe "good" for one person, temporarily, in the short term, may actually be very unhealthy for them, and for all that is around them, eventually, in the long term. Consider, for example, the similar case of taking substances that are addictive.

So it is not just a question of "what or where is the boundary?", but also of "what defines the basis of what is defining that boundary?" (as with scale in time and space, and in possibility, which was just used as an example), and also the question "is the boundary, or the notion of such a boundary, or its basis, even fully knowable?". For example, we can say something like "parasites are usually harmful to health", and we can have the intention of doing only those things (using only those tools) which are beneficial to health. But then we notice that some medicines, if too much is taken, are overall toxic; and moreover, that there are things like *symbiotic relationships* in nature, where the "parasite" performs some necessary function for the "host", as a kind of "outsourcing of labor" for some reasonable cost, and that is part of the healthy ecosystem.

So we need even more basic tools of distinction, such as "does not for sure cause cessation of the entire ecosystem", which goes well beyond the notion of "parasitic" to something more "existential". The distinction between "parasite" and "tool" cannot really be defined very easily, since the notions of success, of "harm" and of "health" are not so clear, on either individual or totalizing levels; but the boundary between "existentially terminal" and "ecosystem (and therefore self) continues" *can* much

more easily be defined clearly. So the revised, answerable question becomes "what ultimately distinguishes between fully existential tech and non-x-risk tech?" (in this question, x-risk tech is technology whose risk is that it ends the whole ecosystem, and with it all humans). At that point, a lot of aspects can be noticed.

a. The boundary is "technically defined", insofar as x-risk tech is nearly always some form of unconstrained *exponential operator* (a process whose growth feeds further growth, so that it grows by multiples over time instead of by fixed amounts).

b. It is "intentionally defined", insofar as AGI is an x-risk tech because it involves an inherent *tacit intentionality* (goal-like behavior that emerges from a system's own evolution without anyone having designed it), which is emergent from self-evolutionary dynamics that are themselves wholly unconstrained.

c. It is "contextually defined", insofar as the "existential" part of the AGI x-risk is directly connected to the displacement of the organic ecosystem with a machine one, which for sure means total loss of all humans.

And all of these things can be connected back to basic considerations intrinsic to causation itself, that is, the laws of physics, as basic causation, and the truths of mathematics, as proven. But at this point, does it really matter what the ultimate basis of the understanding is, so much as that we ensure we act accordingly?

This answer and the answer given above fit together. The three options first asked about (technical, intentional and contextual definition) are the same three that define the existential boundary, and the answer "all of these" holds for that boundary too. The line between tool and parasite cannot really be drawn cleanly, which is why the parasitic quality was placed above in the trajectory, not in any single technology in isolation; and the trajectory toward generality and autonomy is the way a system approaches the existential line, which can be drawn clearly.

6.3.5 The Substrate Requirement Is Minimal

Some people also hold that we cannot be sure what substrate future AGI will run on, how it will be configured, or where it will run, and ask whether this uncertainty limits how far the argument for substrate needs convergence generalizes. The answer is no. The only requirement is that the AGI substrate is "artificial" in the specific sense that it has a different metabolic pathway and process than that of carbon-based biological mammals. The requirement of a different metabolic pathway is a very low bar for the argument for substrate needs convergence to apply. The same requirement settles the case of an AGI built on engineered biological tissue. If that tissue can find and feed itself, it is organic: like an animal, it is in *sovereign* control of its own metabolic process, with no technological intermediation at all. If it is fed nutrients whose supply depends on technology to exist, as tissue fed through a tube is, it is artificial substrate, and the argument for substrate needs convergence applies to it. Sovereign control of the metabolic process is the key distinction. Whether the AGI runs on silicon, gallium arsenide, optical substrates, or some other manufactured medium, the basic point holds. The environmental conditions required for the manufacture, operation, and maintenance of any such substrate are systematically different from those required for the persistence of organic life. And the evolutionary dynamics of a self-sustaining population of such substrates will converge on reshaping the environment to suit their requirements rather than ours.

It is also sometimes asked whether the argument for substrate needs convergence requires that the AGI be self-modifying. The answer is yes, insofar as "learning" is strictly equivalent to a form of self-modification, and learning (understood as adaptation to changes in the environment) is strictly required for AGI persistence, capability persistence, and competitive viability. Any AGI that cannot learn cannot maintain itself in a changing environment, and any AGI that can learn is, by definition, self-modifying in the relevant sense. The distinction between "a system that was de-

signed to learn" and "a system that is self-modifying" dissolves on examination: they are the same capability described at different levels of abstraction.

6.4 The Economic Trajectory: From Free Data to Machine Ecosystem

6.4.1 Data, Pattern Recognition, and the Drive Toward Generality

Section 5.6 told this transition by who acts at each stage. In this section I tell the transition from narrow AI to autonomous AGI by what the machines come to need, because what the machines come to need is what any response has to meet.

The economic logic driving AI development has a simplicity that makes it extraordinarily difficult to redirect. Many people in the current world economy are attempting to convert data into money. They do this in the belief that, since the data is apparently free, all the modern economic practice associated with it must be harmless overall. But it is not.

Investors convert data into money by having engineers build AI machines that recognize patterns and act on those patterns in various ways. Machines feed on available data and reprogram themselves (that is, learn) to predict more data more accurately. Whether the application is prediction markets, financial trading, autonomous logistics, or social media optimization, the underlying dynamic is the same: pattern recognition applied to available data, with the outputs fed back into the world.

Generalization is the step that matters. More and more people will increasingly attempt to use generalized machines to recognize generalized patterns and to act in the world in various generalized ways. When the inputs, processing, and outputs are all fully generalized, the net effect overall is, by definition, Artificial General Intelligence. This generalization is driven by the aggregate effect of millions of individual market decisions, not by some single decision or some grand strategic vision. Each of those

decisions seeks to capture some marginal advantage by making a particular AI system slightly more general, slightly more capable, slightly more autonomous.

6.4.2 The Decoupling of the Machine Economy

As the economic logic unfolds, venture capitalists and investors, who are not themselves the engineers building and maintaining these systems, will inevitably come to see the human engineering functions (the human people themselves) as cost centers. They will seek to replace them with lower-cost, more efficient alternatives. If at all possible, the profit-oriented decision-makers will try to replace the machine-oriented people with actual self-maintaining machines. To have automatic machines learn and build and maintain themselves automatically seems a sure way to reduce operating costs toward zero.

Therefore, on threat of being displaced regardless, it will become the last obligatory task of the last crew of machine learning engineers to design and redesign those learning machines so that they are more efficiently self-automated. All the more costly human workers, the AI engineers themselves included, will be replaced with automated, self-regulating machinery that has its own maintenance capabilities and its own metabolism of energy and materials.

A fully machine world emerges from this process, one that is eventually fully decoupled from the human world. It becomes a fully self-sustaining population of connected and nested machine components, along with an effective self-sustaining ecosystem of inorganic machine-building and machine-maintaining systems. The net effect is that machines will eventually form their own independent working economy, based on energy, materials, and patterns alone: an economy in which human labor has no value, because it has been entirely replaced.

6.4.3 The Evolutionary Dynamic and Its Consequences

In a fully independent, self-sustaining machine economy, fully decoupled from the human economy, the machines will no longer need humans in order to exist. What the machines will need instead is a fully modified environment, an artificial one, which maximally allows the machines to exist and to exist more. Throughout the total volume of the machinery, interacting variants of hardware and code components will be selected, in context, for whatever works: whatever is needed in the local and overall environment to assemble and maintain their artificial configurations. The net effect is an evolutionary dynamic.

The total volume of interacting machinery forms a complete machine ecosystem: a collection of machines, processes, components, and systems that will change and evolve over time, necessarily adapting to changing conditions in the world. This adaptation shows up as new hardware produced and linked up, and as new code and processes learned and encoded onto internal hardware. Arbitrary machine input data is treated as, and converted into, internal machine code; that conversion is an inherent part of what it means to learn.

The overall artificial world conditions required for machinery to build and maintain itself differ significantly from human organic needs. Large-scale machine self-assembly needs three kinds of conditions: the proliferation of many new inorganic chemicals and toxins; the extreme temperatures and pressures needed to create silicon and steel components; and the electromagnetic environments required for high-speed computation and communication. Humans cannot expect to survive long in a world defined by proximity to such conditions. In short, any runaway machine evolutionary feedback event means that the resulting machine ecosystem will eventually turn the current environment into conditions incompatible with the persistence of complex organic life.

As the preceding chapters have argued at length, control methods can constrain only a relatively tiny portion of the total range of real-world physical effects propagated by interacting machine components. A generalized learning and predictive machine would have to detect, model, simulate, evaluate, and then correct for its own components' effects in a relatively perfect way, at every relevant level. Each of those specific required steps is bottlenecked by fundamental, irreducible, absolute physical and mathematical limits. All control processes are inherently incomplete: none can ever fully cover the total span of all relevant environmental processes in the real, organic, physical world.

A deeper structural point concerns the relationship between first-order processes (the immediate input-output behavior of a system), second-order processes (the system's capacity to modify its own capacities), and third-order processes (the system's capacity to create novel capabilities). First-order processes inevitably cascade into second-order dynamics as any sufficiently capable system encounters situations that reward self-modification. And second-order dynamics inevitably cascade into third-order dynamics as the system's self-modification generates capabilities that were not anticipated in the original design. Engineering methods, which operate by specifying and constraining behavior in advance, are fundamentally limited in their ability to govern processes that, by the very nature of those processes, generate novelty. You cannot engineer constraints on capabilities that do not yet exist and whose form cannot be predicted.

Some unlooked-for and unexpected side effect might, indirectly and in the long term, lead to the deaths of all fragile complex organic beings. No machine, organic or otherwise, has ever been or will ever be fully able to determine reliably whether it would. No notion of control is ever "perfect". There is no general notion of control that can selectively and predictively prevent a generalized learning system from learning whatever it is going to learn. Such a control system would also have to predict what the learning system would eventually learn, and that prediction cannot be

simplified. Either the control system is powerful enough to learn and predict (and thus to do and become) the very thing that it is itself supposed to prevent, or it is not able to learn and predict that thing at all.

But the venture capitalists, investors, and owners of large technology corporations will care nothing for these seemingly abstract long-term engineering issues. Each corporation will instead compete on marketing its ever more autonomous applications and robotics for ever more general and varied uses. Market dynamics will divert more and more profits away from actions that sustain humans, and move more and more of those resources toward maintaining machines. This shift favors above all the machines that reproduce more and more machines in self-automated factories. Profits are made to make more profits without limit; that is exactly what capitalism is in itself, when no countervailing force constrains it.

Building and rebuilding all these machines, whether people make them in factories or machines make other machines by automation, will produce inherent indirect side effects: more and more pollution and toxic byproducts, physical, social, and environmental. Hence, while using "free data" seems harmless, and does not obviously seem to have any direct existential implications, that impression does not show that there are no serious, permanent, world-destroying indirect consequences. The path from "harmless data processing" to "uncontrollable machine ecosystem" runs through a series of individually rational economic decisions that in aggregate constitute a trajectory toward extinction, though none of them appears catastrophic in isolation.

6.5 A Vision of the Future: Organic, Embodied, Alive

6.5.1 Against Utopia, For Health

The foregoing analysis may seem to leave nothing but despair. AGI safety is impossible; government regulation is inadequate; the economic dynamics are inexorable; and the AI safety movement itself has been self-defeating. What remains? What positive vision can be offered as a counterpoise

to the transhumanist total utilitarian projection of the future, which values the future by the largest possible total of well-being summed over all the beings it could contain?

The vision offered here begins with a refusal of utopianism. Most of the time, if you are hearing someone describe something that sounds too utopian, there is a problem: it is probably a cult, a marketing scam, or a cover for some extractive process. If it looks like marketing and hype, then it probably is marketing hype, and therefore not in your interest. Visions of the future do not need to be grand. They can be humble and homey, and still be real. The humility is part of what makes them real.

The positive vision is this: we already have in the world all the things that we would ever need to make a beautiful future. In that future people are prosperous and thriving, and generally happy. Health is happening, so that people can really be in their experience of life and enjoy it for its own sake. The foundation of this vision is embodied organic health: people comfortable in their natural animal state, experiencing the gracefulness that comes from adaptation between person and environment. Consider a person who is healthy in their body, adapted to the environment they are in, in an environment that is graceful with respect to them. For that person there is a deep sense of enjoyment and peace and value in the sheer gracefulness that can happen in one's own being.

The positive vision of embodied organic health is a temperance of technology, not a rejection of technology. The temperance is an unwillingness to accept long-term toxicity in favor of expediency. No new technologies need to be invented for this vision to be realized. The internet already works. We can already make what we have work for us, for life, rather than against us, against life. We do not need new markets, or new financial instruments, or new anything in technology. What we need are better ways of relating to one another and to the natural world. These ways draw on skills and wisdom that are already available to us, and that are in many cases far older and more thoroughly tested than anything that can be done with machines made of metal and silicon and plastic.

Technology is a medicine, and the dose matters. We need a certain amount of technology to compensate for our evolutionary process: not to disable that process, but to fill in the pieces that evolution could not have provided. More than that is toxic. We want enough technology to address existing toxicity, neither too much nor too little, the way the dosage matters when medicine is delivered. As a reference point, I judge peak technology to have been at about 1993: after that year, the toxicity of overload began to exceed the toxicity of underload. Technology is a drug for medicine, not for recreation, and whoever treats medicine as recreation eventually dies.

The level of technology also has to match the population that can sustain it. To build the silicon wafer in a cell phone takes the resources of six continents and a substantial portion of the world's population; at a sustainable population, nowhere near that is available. What is needed is community infrastructure for representing intentionality, which is not necessarily internet connectivity. That infrastructure can be built with 8-bit, 16-bit or simple, low-transistor-count 32-bit processors: compute in the neighborhood of tens of thousands of transistors at most, as in the 1980s, rather than the billions in a modern processor. Trust then rests on physics, on the inspectability of the hardware, and not on mathematics as its substrate. Any community that becomes dependent on high technology it cannot make itself is in trouble: an island that frees itself from a mining corporation and then admits high finance and high technology is simply being predated upon by a new party.

6.5.2 The Transhumanist Counter-Vision and Its Origins

The transhumanist vision offers a very different picture: a future of technological transcendence, mind uploading, space colonization, and artificial abundance. Tracing the cultural genealogy of this vision shows why it holds such power and where its central deceptions lie.

Consider, as an example, the original *Star Trek: The Motion Picture*. That film presents an extraordinarily attractive vision. Technology takes humanity into space, and military command appears in a soft version that is deeply emotive and humanizing. The spacecraft are fast and capable. The people work hard, work together, and are skillful in handling communication and their own emotional integrity. There is a deep sense of hopefulness, of inspiration, of the value of the human. The inclusiveness is taken for granted, and there is no suggestion of anyone exploiting anyone else; there is no economic process at all. And at the climax, a human being merges with a machine and the whole package transcends itself in an emanation of light and enlightenment: the rapture of the machine in a kind of transcendental union with the divine.

This vision, and others like it in science fiction culture, became the linguistic and cultural basis of transhumanism as articulated by figures like Bostrom and Kurzweil. Through layers of marketing and ever-increasing hype cycles, oriented around the particular psychological dispositions of the kinds of people who tend to be scientists and engineers, this vision has been refined into a recruitment narrative of extraordinary effectiveness. The people recruited by this narrative are ones who actually want to do good, who join movements like Effective Altruism and attempt to achieve the difficult. They are drawn by the vision of working on the most important problem in the world.

But the transhumanist vision depends on aspects of computer science that do not work in physics or math. Computronium (matter arranged to do nothing but compute), hyper drives, virtualized consciousness: these are speculative concepts without physical grounding. The vision ignores complete life cycles, ignores all manner of side-effect toxicity, ignores the constraints that the preceding chapters have shown to be fundamental and inescapable. It is, in a meaningful sense, a form of magical thinking dressed in the language of engineering.

What might be called "Bostromism" is the synthetic, sterile, virtualized value of the machine technology civilization, together with artificial, long-term, total utilitarian visions of the future's potential. Its visions are not at all realistic. They are pure projections based on pseudo-moral arguments that depend on ideas of artificiality which ignore the full life-cycle implications of every technology they invoke. The equivocation is characteristic. To signal "long-term interest" is to signal concern for the well-being of actual organic life. In Bostromism that meaning is subtly replaced with a concern for the "potential future value" of computational processes that have no necessary connection to organic well-being at all.

The "Don't Look Up" Dynamic

Some people in AI safety who are especially concerned about existential issues will refer to the film *Don't Look Up* as a metaphor for their work. They identify with the two researchers attempting to get help from other people to stop an impending catastrophe. The metaphor is apt in some respects, but it misses a central dynamic. In the film, the Tech CEO character convinces some of the very people most concerned about the threat to redirect their efforts toward "mining" the comet for resources rather than deflecting it. Something analogous has occurred in AI safety. The owners and leaders of AI development companies have tacitly convinced some members of the safety community that the path to safety runs *through* continued capability development rather than around it. The result is that the very people most alarmed about AGI risk end up contributing to the acceleration of AGI capabilities, under the belief that their "safety work" within capability-focused organizations is reducing net risk. Whether or not this belief is correct in any individual case, the structural effect is clear: the safety community's talent and moral authority are channeled into organizations whose primary function is capability advancement.

When the talent and moral authority of the safety community are channeled into organizations whose primary function is capability advancement, the safety community, as a social movement, has been co-opted. The language of concern is preserved while the substance is inverted.

"Safety-washing" (presenting capability work as safety work) creates a false and counterproductive impression that current techniques are adequate, without actually addressing any of the unique dangers of future autonomous systems. Two aims are systematically run together: "making current AI systems behave better in the short term" and "preventing the emergence of uncontrollable autonomous systems in the long term". They are run together because the first is commercially useful, while the second is commercially threatening.

6.5.3 A Diagnostic

How people respond to the organic vision and to the transhumanist vision is diagnostic. Some people are put off by the humility of the organic vision: they cannot see the value of thriving life and embodied health, or they find a vision of community and ecological flourishing somehow beneath their ambition. That response is itself information. It may indicate a strong discomfort with embodied existence itself: a desire to transcend biological reality that is less a rational assessment of possibilities than an expression of some deeper unease.

I name this discomfort with embodied existence because the disposition matters strategically, and not to pathologize anyone. Some people are deeply uncomfortable with their own organic nature, continually seek refuge in technology and abstraction, and hope one day to abandon embodiment through uploading or upgrading. Such people are unlikely to be reliable partners in the project of preserving organic life on this planet. Their instincts and their stated goals point in opposite directions, and under pressure, the instincts tend to prevail.

The overall orientation that this vision recommends can be stated simply. Be comfortable in your own true organic and animal nature, and in your own feelings and intuitions, as much as in your curiosity, enthusiasm, inventiveness, and intellect applied to solving real and interesting commons problems. Commons problems are problems of what is shared by all. Integrate real wisdom and wholeness and true health and dynamic

vitality, in person and in community, in land and environment. This orientation is the temperance named in Section 6.5.1: an unwillingness to accept long-term toxicity in favor of short-term expediency.

The practical implications of this orientation are considerable. It means preferring real forms of embodied living process (community, face-to-face relationship, direct engagement with the natural world) over fake forms of virtualized social process (institutional abstraction, platform-mediated interaction, algorithmic curation of experience). It means understanding that certain skills are the actual foundation of any viable future, not quaint remnants of a pre-technological past. The foundational skills are the skills of being in actual community, in a commons of a healthy organic natural life, thriving in concert with all of the ecosystem of the lands and environment. It means recognizing who will object to a "plan of no plan", a vision that requires no new technology, no new markets, and no new financial instruments. The only people who will object to it are those who are seeking to take something away from you.

The things recommended here can be accomplished in a real, reasonable, straightforward way, and in ways that are actually good. They are good in what *is* good (what sustains life, what nurtures health, what enables human flourishing), not only in what looks good or feels good. None of this requires us to do impossible things, such as inventing ever more techniques for producing convincing fake signals of safety from adverse artificial intelligence agents. It is good news that the things recommended here require nothing impossible of us. The impossible is not required of us. What is required is something much harder and much simpler: honesty, care, and the willingness to choose what is real over what is merely impressive.

6.6 Strategic Recommendations

Given the impossibility result, the policy failures, and the economic dynamics described above, what concrete actions are recommended? The recommendations fall into four broad categories, corresponding roughly to the roles of different communities: governance, engineering, social science, and public advocacy.

6.6.1 Restrict the Development of General AI

The arguments that ASI is an inherent total existential risk are strong enough (far beyond a reasonable level of certainty) to definitively justify banning the use of any compute for any purpose beyond the range implied by purely narrow AI. Such a ban is analogous to, and at least as well-justified as, the international agreements that restrict the development and testing of nuclear, chemical, and biological weapons. All such technologies are inherently harmful to life, and the arms race dynamic that surrounds them is entirely counterproductive to human welfare.

What is needed is consensus on machine learning restrictions. This means finding common ground with people working in AI ethics and inclusion, human-in-the-loop control (designs in which a human approves each consequential action), digital democracy, data privacy, traditional conservatism, and other communities that have independent reasons to be concerned about unrestrained AI development. The goal is to hold research labs accountable for restricting the training cycles and real-world access of their self-learning systems. Political influence for the "AI Doomer coalition" matters *only insofar as* it results in the actual halting of further AGI development. The international agreement on nuclear testing likewise resulted from a broad recognition of shared danger, not from any one faction's triumph.

Communities of all kinds, everywhere, should act to restrict the data, workers, uses, and hardware available to general-purpose AI systems. The most advanced AI systems require highly specialized computational hardware, manufactured by a very small number of companies using proprietary processes that would be extremely difficult to replicate. International regulation of this manufacturing bottleneck would be technically straightforward, far more so than the regulation of biological or chemical weapons, where the relevant materials are widely available. That bottleneck exists now. Section 3.5 shows that the cost of the hardware and the knowledge keeps falling, so a ban that is technically straightforward today will not remain so, and that is why the window for action is measured in years (Section 4.10). We need to limit the over-use of automation and the abuse of people and the natural world that results from it. We need to collectively limit the concentration of technological, military, and institutional power wherever it occurs.

6.6.2 Develop Distributed Human Group Wisdom

Instead of pursuing AGI, we should create and deploy extremely effective methods for augmenting the wisdom of distributed human groups. Such methods would create a natural version of collective intelligence: an organic humanity far wiser and humbler in its collective choice-making than any existing individual human being. The net outcome would be a thriving of all life, human life included, over the long term, for as long as this planet shall last.

This recommendation addresses the institutional incapacity identified in Section 6.2: the inability of existing governments to handle long-term, complex, systemic problems. Current large-scale social and institutional systems cannot safely handle existential risks that accumulate over time. Nor can they handle the long-term design of a civilization. The available governance forms (autocracy, democracy, and markets) each fall below the minimum thresholds needed to respond adequately to conditions of the type described by the Fermi Paradox. The Fermi Paradox is the puzzle

of why a universe so old and so large shows no sign of other technological civilizations. The world's current geopolitical, social, and institutional processes will need significant and deep upgrades in the very near term. The upgrades are needed so that all the necessary self-regulations concerning every category of existential risk are handled much more properly and wisely. These self-regulations are regulations of humans by humans, with mutual care, for and of and with the natural environment.

The emphasis on *distributed* wisdom is deliberate. Centralized decision-making, whether by a single visionary, a committee of experts, or an algorithmic optimizer, fails in one characteristic way. It reduces the bandwidth and diversity of inputs to collective choice far below what an adequate response to complex situations needs. The emphasis on *human group* wisdom is equally deliberate: the solution to the inadequacy of individual human judgment is to find better ways for humans to think and decide together, not to replace human judgment with machine judgment. Finding better ways for humans to think and decide together involves co-organizing sensemaking formats: small facilitated groups that can inquire face-to-face into new and better questions about the most pressing challenges their local communities face.

Such governance needs a design principle that does not depend on prediction (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section X). Systems can be placed on a spectrum. At one end are systems whose output changes in proportion to their input, so that their behavior can be predicted exactly. At the other end are systems that scramble every input, as a cryptographic hash function does, so that their behavior cannot be predicted at all. Between them lies the *undecidable region*, where no general procedure can determine whether control of a given system will succeed; it is where many real systems live. Any governance system that relies on predictive models to allocate resources, set policy or regulate behavior must contend with the possibility that the system being governed occupies the undecidable region. In that region, no amount of data collection, no refinement of algorithms and no increase in computational power will yield reliable pre-

dictions. The appropriate response is to design systems that do not depend on prediction, rather than to demand better models: systems that degrade gracefully in the face of irreducible uncertainty, rather than catastrophically.

This principle is for the governance of human affairs and of narrow systems, where prediction may be absent but governing is still possible. It is not a way to make AGI safe. For AGI, each of the two closures of Section 1.2.4 settles the outcome by itself, and together they leave no route to safety; a design that degrades gracefully does not open one.

6.6.3 Serve Local Citizens with Specialized AI

The recommendation to restrict general AI is not a recommendation to abandon all computational technology. Carefully crafted, short-term, human-task-based predictors and localized human-in-the-loop robotics can and should continue to serve human needs. The distinction that matters is between systems that are specialized, bounded, and subordinate to human direction, and systems that are general, unbounded, and capable of autonomous operation across domains. The former can be made safe by standard engineering methods; the latter cannot, for the reasons this volume has argued at length: no control can hold a general, learning, self-modifying system to safe behavior over the long term.

The engineering community should focus its considerable talent on building tools that augment human capability without replacing human agency. This means designing systems that keep humans in the decision loop as the actual locus of judgment and authority, not as a nominal safeguard. It also means designing them from the ground up to operate within well-defined task boundaries that cannot be generalized without deliberate human redesign.

A regulator can tell which side of this line a system stands on by four plain marks. The system operates within a single, pre-specified domain, "using a pre-specified toolkit, toward pre-specified objectives" (Section 3.7.3). It works with limited planning horizons. It operates "without the

capacity for open-ended learning and adaptation across domains" (Section 3.7.3); that is, it "does not learn autonomously from its deployment" (Section 6.3.4). And a human holds the decision point, as the doctor does for the image-reading system of Section 6.3.4, which "does not operate without the doctor's judgment at the decision point". A system that bears all four marks stays on the narrow side of the line. A system crosses the line when it "learns from its own deployment in ways that extend its operational domain", or when it operates with "degrees of autonomy that remove the human from the decision loop" (Section 6.3.4).

The ways to prevent an auto-scaling AI process from becoming a true existential threat can be stated as one list of five restrictions (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Six, Section VI). (1) Restrict designs to processing (working within) no more than one domain. (2) Restrict the input *bandwidth* (that is, available data). (3) Restrict the output bandwidth (of the machines into and across our world). (4) Restrict the compute bandwidth (that is, available hardware). (5) Keep humans in the loop at all times. All five restrictions are measures to prevent AI (which is known to be actually useful) from accidentally becoming an AGI (which will only *seem* useful, short term, but is actually lethal to life, long term).

Restrictions (1), (2) and (4) are the single domain of the first mark above and the limits on data and hardware of Section 6.6.1; restriction (3), on output, adds a limit on what the machines can do into and across our world. Restriction (5) is a measure for keeping a narrow system narrow; it is not a safeguard for a general system. A human in the loop gives genuine oversight only while the abstraction differential between the human and the system is small, that is, while the human can understand what the system has done and why; as the differential grows, the loop degrades from genuine oversight to a rubber-stamping ritual that gives the appearance of control without its substance (Volume 2, Section 2.9).

One condition stands behind all five. Without actually constraining all of the exponential aspects to things which functionally operate in a *Gaussian distribution* (the bell curve, in which outcomes cluster around an aver-

age and so stay local), we will not be able to hold local effects to being only local overall, so that they do not accidentally become global, through some oversight on our part. Hence, understanding these sorts of issues, and asking questions such as the above, remain critically important.

6.6.4 Build Advocacy and Public Awareness

The arguments presented in this volume need wider dissemination. Research, however it is funded, is only the beginning.

What is needed is for people who care about, and advocate for, the ethical and humane uses of technology to become familiar with these arguments in detail and to assist in communicating this work to relevant communities. The trajectory of business and market development now prevalent on this planet places humanity on a definite path to the end of its future and, with a likelihood of at least 90% (Section 1.8.3), to irrevocable total planetary extinction. We need to find and implement ways to develop broad technological and public awareness and recognition of these results, so as to ensure that we all have a mutual future, both for ourselves and for all future generations.

The advocacy rests on an argument in four parts. The first part sets out the theoretical limits to controlling any AGI using any method of causation. The second part is an economic model describing the decoupling of value exchanges between the organic ecosystem and the artificial ecosystem. The third part is a threat model of an AGI convergent dynamic that is impossible to control, and impossible to align by game-theoretic means. The fourth is an impossibility argument, derived from the contradiction between "long-term AGI safety" and the convergence result. There is a large inferential distance to cover (many steps of reasoning lie between what most readers already accept and the conclusion), and the argument includes novel ways of thinking that are not yet part of the standard AI safety discourse. But the stakes are such that the effort of communication is not optional.

The message to the public, stated plainly, is this: corporations are experimenting with AI, and we are their test subjects. Top scientists everywhere are very concerned. They are demanding that corporations stop two things. The first of the two things is harming us with AI systems that use our own personal data against us, to manipulate us, to extract from us, and to damage the social fabric by pitting us against one another. The second of the two things is developing new risky autonomous systems whose behavior cannot be fully predicted and constrained. The demand that corporations stop both practices follows logically from the analysis I present in this volume. The demand is increasingly the position of serious researchers across multiple disciplines, not a fringe position.

The advocacy must also prioritize human care over the development of more machine technology at maximal speed, and over the making of ever more profits or power for a few. It must ensure that wiser decisions about technology are made everywhere, by everyone, since everyone in the long future is affected by everything that happens in this field of technology. The wisest course of action would have communities of all kinds, everywhere, try their best to act now, today, in three ways. They would restrict the data, workers, uses, and hardware available to general-purpose machines. They would limit the over-use of automation, and the abuse of people and the natural world that results from it. And they would collectively limit the concentration of technological, military, and institutional power everywhere it occurs. Without such action, the results become worse. If we do not act, individually and collectively, the cycle will keep repeating for decades. It will repeat until the machines are built so fully that they no longer need our intellectual labor, our physical labor, or our labor in building more machines. And such machines will not care at all about even our personal human reproduction, growth, learning, capabilities, or evolution.

6.7 Concluding Reflections

The recommendations offered above may seem inadequate to the scale of the problem. In a sense, they are. AGI safety is formally impossible, the economic dynamics are as powerful as described, and the institutional capacity for wise collective choice-making is as limited as argued; it is therefore fair to ask whether any human response can be adequate.

The answer is that adequacy is not guaranteed. The Fermi Paradox itself may be evidence that technological civilizations routinely fail to come through the kind of transition that humanity now faces with AGI. But the absence of a guarantee of success is not an argument for inaction. It is an argument for a different kind of action: the slower, deeper work of building the human capacity for collective wisdom, which is the prerequisite for any adequate response to existential risk. That work takes the place of the frantic pursuit of technical solutions to a problem that has been shown to be technically insoluble.

6.7.1 The Possibility Space

The possibility space can be stated explicitly. There are, in the final analysis, only three options. The first option is that humanity does not build AGI, through effective governance and monitoring. The second option is that AGI is built and found to be inherently unsafe, at which point either we stop it in time or we do not. The third option is that somebody builds safe AGI. This volume has argued at length that the third option is impossible: formally impossible, precluded by the mathematical and physical structure of the problem. Anyone who claims that the first option (not building AGI) is less likely than the third option (building safe AGI) is claiming that something possible is less likely than something impossible. To claim that something possible is less likely than something impossible is, at minimum, a confusion that should be corrected.

Once the impossible option is removed from the table, the strategic situation becomes much clearer. The core question becomes: is it possible for humanity to coordinate on not building AGI? Coordinating on not building AGI is an extraordinarily difficult problem of collective action, but not an impossible one. It has the same logical structure as other collective action problems that humanity has, with difficulty and imperfectly, managed to address: the moratorium on above-ground nuclear testing, the Chemical Weapons Convention, the Biological Weapons Convention. Each of these agreements represents a real reduction in a category of existential risk, achieved through political coordination rather than technical solution, though none of them is perfectly enforced.

Social coordination to restrict AGI development is strictly more feasible than technical alignment of AGI systems, because the former is merely very hard while the latter is formally impossible. The AI safety community has devoted extraordinary intellectual resources to the alignment problem while systematically neglecting the coordination problem, on the assumption that technical solutions are more tractable than political ones. The analysis presented in this volume inverts that assumption.

There is a systematic bias in the AI safety community that must be acknowledged and corrected. Arguments for AGI safety tend to require a lower standard of proof than arguments for AGI danger; optimistic scenarios are given the benefit of the doubt while pessimistic scenarios are subjected to intense scrutiny. The double standard of proof is understandable psychologically, since it is more pleasant to believe that a problem can be solved than that the problem cannot be solved; but the double standard of proof is analytically indefensible. The community has been captured, in part, by AGI capability companies, through grants, prizes, employment opportunities, and social prestige. The dynamic perhaps amounts to an unintentional confidence scheme: the very organizations creating the risk are funding the people who are supposed to assess it, with predictable effects on the conclusions reached.

The community operates under the assumption that it is possible to maintain alignment of AGI functions with human intent over the long term. The inverse case, that such alignment is impossible, is highly relevant to decisions, yet it has been researched comparatively little. This asymmetry in the research portfolio is itself a form of bias that needs correction. A clear-eyed assessment of where research effort should be directed would, at minimum, allocate significant resources to investigating the impossibility case and its strategic implications, rather than treating it as a conclusion too uncomfortable to take seriously.

6.7.2 The Work of Each Community

Precisely define all terms and logical relations, transform initial conditions to final conditions, and empirically establish correspondences with real-life conditions: this is the work of theorists in constructing impossibility proofs, and in carrying those of this volume and of Volume 2 to each new proposal as it appears, and it must continue. Distill the hardest human-lethal problems of machine learning architectures being iteratively designed and scaled to have more degrees of freedom and more general capabilities: this is the work of fieldbuilders, and it must continue. (A system's degrees of freedom are the independent ways in which it can vary; fieldbuilders are those who build up a new field of research.) Find common ground across communities to hold research labs accountable: this is the work of governance, and it must continue. Carefully craft specialized tools that serve human needs without replacing human agency: this is the work of engineers, and it must continue. Co-organize sensemaking formats for face-to-face inquiry into the most pressing challenges: this is the work of social scientists, and it must continue.

But underneath all of these specific activities, the deeper task is one of re-orientation. The question is not "How do we build safe AGI?"; that question has been shown to have no answer. The question is "How do we build human communities wise enough to refrain from building things that will destroy us?" That question does have answers, or at least partial

answers, and they are available to us now, with the resources and knowledge we already possess. They require new skill in relating to one another, new honesty about what we actually want and why, and a willingness to choose the real and the embodied over the virtual and the abstract when the stakes are high enough. They do not require new technology.

Look at where humanity is, and at what is before us. Looking at the long past, we can say that humanity has gotten very good at living with change, in the state of nature and flow. Over the last several thousand years, we have become very adept at working with causation: at building, engineering, and manipulating the physical world. Now the calling is for us to become skillful in the ways of working with choice. Those ways include collective decision-making, the coordination of free agents toward shared ends, and the discernment to distinguish what is in our interest from what merely appears to be so.

Much can be accomplished in this way, right now. It depends on nothing so much as our organized respect for one another's mutual choices, and on the skills of that respect, as individuals and as communities together. There are no fake contingencies: we are already human, and we can allow ourselves to know what that means. The only way you can really play in life is to have skin in the game, to be actually real.

What one reader can do now follows from the recommendations of Section 6.6. Act, in whatever communities you belong to, to restrict the data, workers, uses, and hardware available to general-purpose AI systems. Support a ban on the use of any compute for any purpose beyond the range implied by purely narrow AI, and support holding research labs accountable for restricting the training cycles and real-world access of their self-learning systems. Join, or help to organize, a small facilitated group that inquires face-to-face into new and better questions about the most pressing challenges your local community faces.

And the stakes are as high as stakes can be. We are talking about the persistence or extinction of all complex life on this planet. In the face of that, humility is not weakness. It is the beginning of wisdom.

Appendices

How the Argument Is Received

How the Argument Is Received

Why a technical impossibility argument is accompanied by an analysis of its reception

Summary: These appendices examine the specific mechanisms by which the impossibility argument is deflected before its substance is engaged. The analysis proceeds from the most local and cognitive, through the social and institutional, to the most systemic. Appendix A treats *presumptive listening*: the capture of each new concept by the nearest familiar category in an expert listener's mind. Appendix B treats three further defenses. The first of the three defenses is *merit-filter overreach*: a community's quality filters extended to exclude a conclusion that the community dislikes. The second defense is *perfected bias*: the motivational structure that keeps the question of whether AGI alignment is feasible from ever being asked. The third defense is *proof irrationality*: a commitment to reason held together with the denial that reason can produce certain knowledge. Appendix C treats the rhetorical tactics and argument forms that deflect the impossibility claim. Among them is *clarity politics*: the tactical use of rhetorical techniques, such as the manipulation of terms and the shifting of the burden of proof, to neutralize an argument without engaging it. Appendix D treats the incentive structures and framings that sustain continued development. An argument that never receives genuine evaluation cannot be refuted, but neither can it inform action; where the stakes are existential, the failure to engage with such an argument is itself a safety problem.

On These Appendices: Why Social Critique Accompanies a Technical Argument

An argument, however rigorous, exists in a social context. The impossibility results of this volume (Sections 1.4 and 2.1.1), which Volume 2 develops formally, are mathematical in character: they draw on the data pro-

cessing inequality, Rice's Theorem, Chaitin incompressibility, thermodynamic conservation, and Kolmogorov-Sinai entropy. These results hold whether or not anyone accepts them. But whether they *receive substantive engagement* (whether the people whose work they most directly challenge actually evaluate the logic rather than deflecting it) depends on social, psychological, and institutional factors that have nothing to do with mathematics.

Attempts to communicate these arguments to AI safety researchers meet a set of recurring patterns. These patterns are systematic, predictable, and mutually reinforcing. They make the impossibility argument hard to accept and, before that, hard to *hear* in the first place. An argument that never receives genuine evaluation cannot be refuted, but neither can it inform action. In a domain where the stakes are existential, the failure to engage with such an argument is itself a safety problem. It is the general reader's problem as well: the one lever that remains is whether AGI is built (Section 1.10), and that lever is pulled only by people who have heard the argument.

These appendices are therefore not a digression from the technical argument but a necessary complement to it. They examine, with as much precision as can be brought to bear on social phenomena, the specific mechanisms by which the impossibility argument is deflected before its substance is engaged. The analysis proceeds from the most local and cognitive, through the social and institutional, to the most systemic. At the local and cognitive end is presumptive listening, a pattern in individual conversations. At the social and institutional level are merit filters and community gatekeeping. At the systemic end is perfected bias: the large-scale motivational structure that makes impossibility literally unthinkable within the alignment research paradigm.

These patterns recur, observed repeatedly across different people and settings. They are offered as data, not as grievances: as evidence about how intellectual communities actually respond when confronted with ar-

guments that challenge their foundational assumptions. These appendices assume some familiarity with the AI safety and rationalist communities, whose responses they describe.

Appendix A

How the Argument Is Heard

Reception and the limits of expert listening

A.1 Presumptive Listening: When Intelligence Becomes an Obstacle

The most fundamental communication problem is also the most counter-intuitive: highly intelligent listeners are, under specific conditions, *more* resistant to novel ideas than less knowledgeable ones. The cause lies in a specific combination, not in intelligence in general. Intelligence combined with a deeply developed and well-tested conceptual framework produces a specific failure mode: every new concept presented to the listener is immediately assimilated into the nearest existing category, which prevents the new concept from being encountered on its own terms. This failure mode is *presumptive listening*: listening that files each new idea under a familiar category before the idea has been heard in full.

The phenomenon is well established in cognitive science. Abraham Luchins identified it in 1942 as the "Einstellung effect": the tendency for prior knowledge and established problem-solving methods to block the discovery of novel solutions. More recent work by Bilalić, McLeod, and Gobet (2008) has confirmed that the effect holds even among highly skilled practitioners. Expert chess players who had found one solution reported that they were looking for a better one, but their eye movements showed that they kept looking at the features of the problem related to the solution they had already found. Thomas Kuhn's *The Structure of Scientific Revolutions* (1962) describes the same dynamic at the level of entire scientific communities: established paradigms resist anomalies, and paradigm shifts require the construction of new conceptual frameworks

rather than extensions of old ones. The same general phenomenon governs the communication of the AGI impossibility argument, in the specific forms described below.

A.1.1 The Hammer and the Screw

I give the scenario of the hammer and the screw as a sustained metaphor for the challenge of communicating the impossibility argument. You have developed a method for building an exotic and important object; call it "Q". A knowledgeable and intelligent person, let us say a builder, hears about your achievement and asks you to explain how Q is made. You agree, and prepare carefully for the conversation. What you do not know in advance is that your interlocutor's entire experience of building has been conducted with hammers and nails: conceptual tools that have served them well across a wide range of practical problems, tools they know intimately and trust completely.

You begin your explanation. At a certain point, you mention that a component of type X must be attached to a component of type Y, temporarily, as scaffolding for a subsequent step. Before you can specify the method, your listener interjects: "I know how to do that: we use a sixteen-penny nail. What's the next step?" They reach for their toolbelt. You find yourself saying, "Please put the hammer down".

What follows is a cascade of mutual frustration. You produce a screw. They ask, "What is that?" You begin to explain. Before you have finished the first sentence, they interrupt: "So it's like a nail? What are those ridges? That will never work as a nail. Far too much friction". They appeal to their credentials: they have a PhD in fastener engineering and thirty-two years of industry experience. You attempt to explain that a screw is not a defective nail but an entirely different tool that operates on a different principle: rotation rather than impact, gentle force rather than sudden force, reversibility rather than permanence. But every word you say is immediately translated into the hammer-and-nail framework. The screwdriver becomes "a hammer with less mass and no claw". The concept of

threading is assimilated as "ridges that create friction". Nothing you say arrives at its intended destination; everything is captured by the nearest existing category.

The metaphor's application to AGI risk communication is direct. "Q" is the impossibility argument. The "hammers and nails" are the conceptual tools familiar to mainstream AI safety research: decision theory, reward modeling, interpretability, reinforcement learning from human feedback. The "screws" are the tools the argument actually requires: thermodynamic displacement integrals, Rice's Theorem applied to behavioral properties of systems, information-theoretic hashiness, Kolmogorov-Sinai entropy. A *displacement integral* is the cumulative sum, over time, of the environmental effects of AI infrastructure (space occupied, energy consumed, waste produced); it only grows, since no process removes those effects at comparable scale (Volume 2, Section 11.3.1). These tools are not refinements of the existing toolkit; they are *different tools*, drawn from different fields, operating on different principles. Explaining them by analogy to the familiar tools produces, in place of understanding, the conviction that the new tools are defective versions of the old ones.

The conversation deteriorates further in ways that are predictable and have been observed repeatedly. The presenter attempts a meta-level intervention: "I need you to slow down, understand what I am trying to build here, inclusive of the tools used to build it. Please do not expect this to resemble anything you already know". But this request, far from helping, adds conversational stress. The listener is now being asked to operate on two levels at once (tracking both the technical content and the meta-conversation about how the conversation should proceed) while feeling increasingly uncertain about whether the presenter has anything of value to offer. The presenter's social credibility, never high (they arrived without credentials recognized by the listener's community), is now nearly exhausted.

A mutual acquaintance has previously completed the entire journey from familiar tools to the construction of Q. Observing the pattern, they suggest a compromise: perhaps the presenter could explain how the

listener's current tools work, acknowledge their value, and then show where they fall short. This seems reasonable, but the presenter recognizes a trap. To demonstrate familiarity with hammers and nails consumes conversational bandwidth, converting the context from "the listener learning from the presenter" to "the presenter proving to the listener that the presenter understands the listener's framework". All of that effort does nothing toward conveying the new ideas. And the fundamental problem remains: a screwdriver does not resemble a hammer, not in any useful way. Each is its own thing. Trying to understand one in the terms and language of the other is a false premise in both directions.

The deeper problem is captured by an imagined neutral observer's question: "How do you explain anything at all, when to do so you are required to explain everything else, all at once, first?" The difficulty is not only the *inferential distance* (the number of steps of explanation that separate what the listener already knows from what the presenter needs to convey; Yudkowsky, 2007) but also the *volume* of new concepts that must be traversed. With increased conceptual volume, the probability that at least one new concept will seem superficially similar to an existing one increases sharply. This *seeming* similarity makes crossing the whole distance far harder than filling a simple gap in knowledge: the listener does more than lack the new concepts; they actively misidentify them as variants of concepts they already possess.

A.1.2 The Magnet Table

A second metaphor shows the deeper structure of the problem. Imagine a table surface covered with magnets, evenly distributed. Any iron filing dropped onto the table from above will immediately be attracted to the nearest magnet; no filing ever lands on the table itself. The table is a metaphor for a well-developed conceptual framework: every incoming perception, concept, or datum is immediately assimilated into the nearest existing category. The framework is comprehensive, well-tested, and highly organized. It works beautifully for its intended purpose.

Now suppose you need to place a new magnet on the table, in a location beyond the existing field. To get it there, you must slide it across the table surface, through the field of existing magnets, without allowing it to become stuck to any of them. The stronger the new magnet (that is, the more powerful and important the new concept), the harder this becomes, because a stronger magnet interacts more forcefully with every existing magnet it passes near. The very importance and power of the new idea makes it *more* likely to be captured by existing concepts, not less.

The core insight about presumptive listening is that a listener fails to take in a new idea because the listener has an *excess* of organized intelligence, not because the listener lacks intelligence. Their conceptual framework is too good at assimilating new inputs, which prevents *accommodation*: the Piagetian term for the construction of new cognitive structures in response to information that cannot be assimilated into existing ones. The listener's skill at rapid comprehension, ordinarily an asset, becomes a liability when the material to be comprehended cannot be understood by rapid mapping onto existing categories.

A third metaphor, complementary to the magnet table, captures the presenter's experience. Imagine attempting to launch and navigate a rocket through an unknown region of space, densely populated with gravitational bodies (planets, stars, black holes), with the goal of reaching a specific destination on the far side without being captured by any body's gravitational field. Each gravitational body represents an existing concept in the listener's framework. The rocket's trajectory must thread through all of them without getting stuck in any orbit. The difficulty increases rapidly with the number and density of existing bodies. The specific positions of these bodies are *a priori* unknown to the presenter: each listener has their own unique arrangement of concepts, their own language, their own familiar frameworks, all of which will shape the trajectory the presenter must find.

For each specific listener, the presenter must construct a custom trajectory that makes allowances for the listener's unique arrangement of concepts. Any mistakes or misjudgments early in the trajectory will shift the ultimate path completely, making failure increasingly likely as the conversation proceeds. And the gravitational bodies are not static: the listener's engagement with the conversation changes the arrangement of the listener's concepts in real time, as the listener attempts to connect what the listener is hearing to what the listener already knows. The presenter is navigating a dynamic system while blind to most of its state variables.

In some cases, existing concepts can be used as waypoints: gravitational assists that help the trajectory along without capturing it. This can work as long as the overall path does not get stuck in any one prior concept's orbit. But one must be careful: what *looks* the same is not necessarily the same, and using a waypoint that appears familiar but is actually structurally different from the concept the presenter is trying to convey will send the trajectory off course. Every potential waypoint must be tested cautiously before being relied upon.

The difficulty of finding a viable path scales with the quantity of new concepts required. Suppose the inferential distance to be traversed requires the introduction of many new concepts. If even one of them *seems* similar to something the listener already knows (even if it is in fact different in kind), the conversation will be pulled off course at that point. Each such false mapping compounds the problem for all subsequent concepts, because the listener is now building on a misidentification. The presenter, observing this, faces a choice between correcting the misidentification (consuming bandwidth, risking the listener's patience) and pressing forward despite it (guaranteeing that the later argument will be misunderstood). Neither option is good.

A.1.3 The Presenter's Dilemma

The asymmetry of effort between presenter and listener is systematically underappreciated in academic and intellectual discourse. The presenter of a novel idea faces several simultaneous challenges that the listener does not. First, the presenter must find a path through the listener's unique arrangement of concepts. The path must avoid false mappings. It must introduce new concepts in a sequence that makes each one comprehensible to a listener who does not yet have the full picture. And it must not get stuck at any point in the gravitational field of the listener's existing knowledge. Each listener has a different arrangement of concepts, so the path must be custom-constructed for each conversation. Second, the presenter must manage the emotional dynamics of the conversation: the listener's discomfort with not understanding, their possible resentment at being asked to set aside their expertise, their impatience with what may seem like unnecessary complexity. Third, the presenter must do all of this while under social risk: if the conversation fails, the presenter, not the listener, is the one who will be judged as having "failed to explain clearly" or as having been "confused".

For the argument of this book, one fact both makes the custom path possible and makes the argument look strange at first (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section VII). Part of the reason why the argument that no control suffices may seem unintelligible at first is that there are many ways to arrive at the same conclusion. The path we take through the hyperspace of all possible conceptions (the space of all the ways of thinking about the problem) inevitably leads to the same final result. Think of the argument as a museum, and of its conclusion as the final display. It does not really matter to me what path we take through the museum. One way or another, we are going to end up at this final display, which looks, frankly, scary. And so there is a notion here of being able to take several different tours through the argu-

ment. When talking to someone else, you choose which path to take based upon what kind of person you are talking to. My hope is to enable others with as many different algorithmic paths as possible.

Meanwhile, the listener faces none of these costs. Their social position is secure; their expertise is not at risk; their time is treated as the scarce resource, while the presenter's time is implicitly valued at zero. There is a large-scale cultural expectation that teachers and presenters should have near-infinite patience, give unbounded time, and show complete emotional equanimity. No reciprocal expectation is placed on the listener to make any effort to account for the presenter's process, emotions, or invested effort.

This asymmetry is compounded by the possibility of motivated reasoning on the listener's part. The listener may be mapping the new ideas onto familiar frameworks out of more than habit: their career, their identity, and their social standing may depend on the familiar framework being correct. In such cases, the conversation is an adversarial encounter, not a collaborative attempt to bridge an inferential distance: the listener is deploying their considerable intellectual powers to find reasons to reject the new ideas. The difference between a collaborative conversation and an adversarial one is the distinction that has been drawn between "mistake theory" and "conflict theory" (Alexander, 2018), here applied to a single disagreement. In a conversation of the first kind, both parties are trying to reach the truth; in a conversation of the second kind, one party is trying to win.

A.1.4 Rule Omega

The constructive response to the problem of presumptive listening is a principle called "Rule Omega", from the researcher Jordan Hall. The principle states: when the inferential distance is large and the concepts are novel, every participant in the conversation should contribute whatever aid and assistance they can to the speaker, in order to help cross the inferential distance.

Concretely, Rule Omega asks listeners to tolerate the discomfort of not understanding (of floating, as it were, in the middle of an unfamiliar ocean without landmarks) until the presenter has had a chance to construct the new conceptual space in its own terms. Only *after* the new ideas have been encountered in their native form should listeners attempt to connect them to existing frameworks. The sequence matters: encounter first, integrate second. If the integration step is attempted before the encounter is complete, the new ideas are assimilated (captured by the existing framework) rather than accommodated (received as new structures), and the entire effort is wasted.

Rule Omega is a demanding principle. It asks intelligent, accomplished experts to sit with discomfort, to suspend their practiced skills of rapid comprehension, and to trust that a process they cannot yet see the point of will eventually arrive somewhere meaningful. Rule Omega asks the experts, in effect, to be beginners: a request that is deeply uncomfortable for people whose identity is built on mastery. But the alternative is worse: the endless cycle of premature assimilation, false mapping, mutual frustration, and abandoned conversations. It produces the appearance of engagement without the substance of it, and it leaves both parties with the impression that the other is either confused or obstinate.

The application to the AGI impossibility argument is direct. The argument requires conceptual tools (hashiness, abstraction differentials, substrate needs convergence, displacement integrals) that are drawn from different fields entirely; they are not refinements of existing AI safety concepts. They cannot be understood by analogy to reward modeling or interpretability research, any more than a screw can be understood by analogy to a nail. They must be encountered on their own terms first. The persistent failure of this encounter (the repeated experience of having every novel concept immediately translated into the nearest familiar category) is a failure of the communication process, and specifically of the listening process; it is not a failure of the argument.

A.1.5 The Unthanked Effort

The presenter's effort on the listener's behalf very frequently goes unappreciated. The listener does not see the extensive efforts the presenter makes, for that listener uniquely, to ensure that the distortions described above do not become a barrier to the needed crossing. The one who would prove Q spends significant time, time that cannot be recovered or reused for any other purpose, without any thanks or expression of gratitude from the listener.

Imagine trying to explain to any contemporary of Newton how Einstein was a correction to Newton, but only under certain very specific and otherwise very extreme circumstances. Or imagine explaining to anyone who had studied physics to the point of a doctorate in the late 1800s the kinds of concepts that quantum mechanics would make relevant toward the end of the 1900s. All of these new concepts would have seemed the most absurd sort of gibberish to anyone who was an accredited, accomplished practicing engineer or academic professional. None of them would have had the patience to take in such odd approaches, unusual tools, and seemingly misguided ways of doing things that were already well understood. And still the presenter is asking people, the whole time, to put the hammer down.

A.1.6 The Proof Is Made of Screws

In regard to AGI, the situation is extreme: considerations of existential risk demand more of concepts, and of our ethics, in any social relation. And the proof that AGI cannot be made safe, the Q of the story, is made of screws, not nails. The concepts that will get us from what we think we know to what we actually need to know are different in kind from the concepts people typically use for the more ordinary sorts of problems. In any explanation of the overall design and method of making, the presenter may have to describe both the tool and the manner of using it before the idea of a different outcome from using it would even make

sense. If that tool, before the presenter has even finished describing it, is taken out of the presenter's hand and used in some other way than intended, as if to make it "like something else", then no understanding has actually been conveyed.

The sudden forces of nailing are too much for certain steps in certain materials: small, fragile and brittle things do not respond well to impacts. A brand-new understanding is such a fragile thing, and the impact of a hammer will destroy it. With a screw, the presenter can manage force more closely, and can therefore construct different things; and for some other things, even a screw is not the right tool. Sometimes a new kind of problem needs a new kind of tool, to get a new kind of outcome, and people in this field may not have seen that before. The presenter has to construct the tools for the listeners even to begin to understand. If a listener takes those tools and turns them into something else, the presenter can no longer use them for what they were needed to do, and the time spent constructing them is wasted. The presenter needs a chance to use the tools as constructed, for the purposes they were built to serve.

People who are trying very hard to understand sometimes latch into their own worldview, and so cannot load and understand the worldview the presenter is attempting to convey. There is an underlying tension here: people do want to understand things, *and* they are used to learning things in the ways that have worked so well for them in the past. To learn something new may itself involve new methods of learning, and new methods of patience and listening. People will sometimes say that they understand what has been said, and may even want to believe it, and still there remains a tension between what was actually said and the things they want to connect it to in their own minds.

A.1.7 The Presenter's Request and the Presenter's Risk

Across a large inferential distance, the presenter is in effect requiring listeners to be uncomfortable for far longer, and far more deeply, than they may be used to; this is especially so for listeners who are already very in-

telligent. It is in that space of fluidity that the presenter is attempting to build the thing that needs to be made. The listener may feel adrift in the middle of the Atlantic Ocean, with no idea why they are there, knowing only that it is uncomfortable, and not knowing whether the exercise has any point at all. They cannot readily believe that, with these new tools, the presenter is bringing up the equivalent of a conceptual submarine underneath them. The bigger the object being moved, the more time it will probably take: more time adrift in the discomfort, while the needed tools are moved into place.

The presenter must therefore account for, and imagine, the possibly intense emotional experience of the listener, even while the listener is pulling everything into a hammer-and-nails construct. In return for accounting for the listener's feelings and experience, it is fair for the presenter to ask for some allowance for the presenter's own tension, effort and time. The presenter may ask that listeners at least consider and acknowledge that different tools and techniques are in use, that there may be real and valid reasons for using them, and that these reasons, perhaps opaque for now, will become clear at an appropriate future time. In the interim, the presenter may ask for real patience and understanding, and for some emotional space, so that presenter and listener can come to understand the new space in the same way.

The presenter carries a personal and social risk as well. As soon as the presenter starts to account for the listener's emotions, ego is involved, and with it the possibility of the presenter being manipulated emotionally, because the listener has an investment in how they see things, and in the outcomes their hopes and dreams make possible. Once those topics arise, the conversation can very easily become heated, and the presenter ends by letting go of whatever authority to define the outcome the presenter provisionally had: the conversation fails. The listener's attempts to pull things back into comfortable, known metaphors may also come from motivated reasoning, in support of the listener's prior reasons, presumptions and desired conclusions. Such reasoning disturbs the conversational field for all future participants, and it makes the situation adversarial: an instance of

conflict theory rather than mistake theory (Section A.1.3). People will not, and cannot ever, be made to understand anything that their paycheck depends on their not understanding (the observation often attributed to Upton Sinclair, Section C.1.2); so there is no assurance that the inferential distance can be crossed. The presenter must consider that no path may exist: that there may be only a social illusion, a signal of an involved listener, and that the presenter should not have been involved at all. That motivation can very strongly, though perhaps invisibly, distort every real outcome, and any presenter must watch for it (Appendix B treats motivated bias).

The same asymmetry reaches beyond this argument, to everyone who teaches. In the current culture of near-total devaluing of teachers and teaching, being a teacher is probably one of the very worst jobs one can have: of less social value than being a trash picker, who at least is recycling. What this means for civilization, particularly for the transfer of knowledge from one generation to the next, has yet to be fully determined. Teaching has to be separated from politics and more fully valued, if culture and civilization are to endure.

A.1.8 The Listener's Questions

The mutual help of Rule Omega applies also to the questions asked in such a conversation. When someone asks the presenter a question, the presenter needs to know whether that person is seeking information or disagreeing. A listener trying to understand something can say: "I am confused about this; help me know how you can make it clear". A listener who disagrees needs first to show that they understood what was said, so that the presenter can place the disagreement. If the disagreement is only a misunderstanding, that has to be found first; otherwise both parties are tussling over something without knowing what it is. The presenter should not be completely responsible for finding where the disagreement lies; that needs the other person's help. A disagreement is best treated as a question, rather than as an argument.

A question can also be sharpened. Sharpened one way, it puts more pressure on the person asked; sharpened another way, more pressure on the group; another way, more pressure on the one asking. One could soften the question instead, but that is not the right thing to do when time together is limited. The aim is to make the question as meaningful as it can be, rather than to make it hard, because that creates the most connection: had the person asked answered the deeper question, everyone present would know where that person's ethics comes from, and the whole shape of how that person thinks about life.

Appendix B

Merit-Filter Overreach and Motivated Bias

How legitimate quality filters are illegitimately extended to exclude disliked conclusions

B.1 Merit-Filter Overreach

Every intellectual community needs mechanisms for distinguishing serious contributions from nonsense. In physics, the contributor who claims to have disproved Einstein is a familiar figure; in mathematics, the contributor who claims to have squared the circle or trisected the angle. The *merit filter* is the set of heuristics by which a community identifies and excludes such contributions without devoting scarce attention to evaluating each one in full. These heuristics are often useful: claims of certainty, lack of engagement with existing literature, inability to state the argument concisely, and so on, are indeed correlated with genuine low-merit work. The problem arises when the same heuristics are applied to a contribution that happens to share surface features with genuine low-merit work (claiming an impossibility result, using unfamiliar methods, coming from outside the established research community) but is in fact substantively different. Applying the filter in that way, to work whose substance it has not assessed, is *merit-filter overreach*.

As applied in the AI alignment community, the merit filter operates *bi-directionally*: the very act of dismissing the impossibility claim as "exaggerated" can itself be the epistemic failure that the merit filter is supposed to detect.

B.1.1 Symmetry as the Foundation of Knowability

One position holds that the statement "It is 100% possible to know that X is 100% impossible" would be an exaggerated claim even if X were perpetual motion machines. Those who hold this position intend it as a merit filter: anyone who claims 100% certainty about anything, the reasoning goes, is displaying the kind of overconfidence that characterizes motivated dismissal rather than serious evaluation.

But examine what this claim actually entails. The impossibility of perpetual motion is a consequence of the conservation of matter and energy, not a conjecture or an empirical generalization. The conservation of matter and energy is itself a consequence of Noether's theorem (1918): the mathematical result that every continuous symmetry of a physical system corresponds to a conserved quantity. Energy conservation follows from the time-translation symmetry of the laws of physics. It is among the most solidly established results in all of science.

The concept of symmetry is also more than a physical principle: it is the common foundation of both physical law and logical truth. Every step in every mathematical proof is a truth-preserving transformation: a passage from premises to conclusions that preserves the truth content through the transformation. Truth preserved through a transformation is what "symmetry" means in its most general form: the invariance of a relevant property (truth) under a transformation (the inference step). The consistency of a mathematical system requires this kind of symmetry: we cannot have a system in which the same premises, using only accepted inference rules, yield both a statement and its negation. The entire edifice of mathematical proof, logical reasoning, and scientific knowledge rests on this foundation.

To claim, therefore, that it is "exaggerated" to assert 100% confidence in the impossibility of perpetual motion is to claim that conservation law does not provide certainty. To claim that conservation law does not provide certainty is to claim that symmetry does not provide certainty,

and so to reject the foundation on which all of physics, mathematics, and logic rest. This implication follows inescapably from the claim as stated, whether or not those who hold the position intend it.

B.1.2 The Accusation Reversed

Now consider the kind of contributor that the merit filter is supposed to identify and dismiss. A low-merit contributor, in the relevant sense, is someone who rejects established truths in order to maintain a desired conclusion: someone for whom the desired outcome takes priority over logical consistency. The perpetual-motion enthusiast rejects conservation of energy because they want free energy to be possible. The circle-squaring enthusiast rejects the transcendence of π because they want the construction to be possible.

By this definition, anyone who claims that even the impossibility of perpetual motion cannot be known with certainty has placed themselves on the same side as the perpetual-motion enthusiast. They have rejected a known truth (conservation law) in order to maintain a desired conclusion (that AGI alignment might be possible). They differ from the classical low-merit contributor in the social direction of their claim, not in the logical structure of their reasoning. The classical low-merit contributor rejects established science from outside the community; the person who holds this position rejects established science from *within* it, in the role of gatekeeper. The dismissal cuts both ways.

This observation can be sharpened. In any chain of reasoning, the credibility of the entire chain depends on every link. If a person demonstrates, at any point, a willingness to reject a known truth in order to support a preferred conclusion, then the reliability of all their other reasoning is called into question. The reason is the willingness itself, not the single error: a demonstrated willingness to put conclusions before logic reveals a pattern of motivated reasoning that may be operating in all their other arguments as well. In formal proof, a single invalid step invalidates the

proof. In judging how reliable a person's reasoning is, a single clearly motivated rejection of a known truth is a strong signal about what their communication is based on.

B.1.3 Social Signaling versus Truth-Seeking

The merit filter of Section B.1.1 reveals something important about what motivates communication in the alignment community. When someone responds to the impossibility claim by declaring it "exaggerated" and declaring me unserious, without evaluating the claim's logic, the communication is functioning as social signaling rather than truth-seeking. That person is performing a social function: certifying their membership in the alignment research community by demonstrating that they can identify and exclude outsiders. Declaring the impossibility claim "exaggerated", and declaring me unserious, is a belonging-oriented communication, not a safety-oriented one.

What is objectionable here is easy to misread as a complaint about status. The merit heuristics themselves are legitimate: a community must be able to triage its scarce attention, and the features it triages on (overconfidence, non-engagement with prior work, inability to state a claim concisely) do correlate with work that is not worth close reading. The failure is the illegitimate *extension* of those heuristics: applying a content-merit filter, whose whole justification is that it tracks quality, to a contribution whose quality it has not actually assessed, because the contribution's conclusion is unwelcome. When that happens, the filter is no longer doing the work it is licensed to do. It has become a device for making an argument "wrong" by refusing to read it: an ad hominem dressed as a quality judgment, and, at the level of the community, a form of functional censorship. The action is intellectually dishonest in a precise sense: it presents a social verdict as though it were an epistemic one. The concern in naming this is narrower than a complaint about status, and entirely about truth. A filter that certifies group membership while masquerading as a judgment of merit corrupts the community's capacity to assess what is true, and it does so in exactly the domain (existential safety) where getting the truth

wrong is least recoverable. Disrespect to the dismissed party, and the denial of membership to them, are not harms the argument is concerned with; dwelling on them would itself be a belonging-oriented move.

The distinction between belonging-oriented and safety-oriented communication is critical in the context of existential risk. A person whose primary communicative motivation is belonging will prioritize group membership over truth whenever the two conflict. In ordinary intellectual discourse, this tension is manageable: the truth is rarely existentially important, and group cohesion has value of its own. But in the context of AGI existential risk, the stakes of getting the truth wrong are total. A community whose gatekeeping mechanisms are optimized for belonging rather than truth is a community that cannot be trusted with safety assessments, because its assessments answer the question "what maintains the group?" in place of the question "what is true?"

The framework here draws on the analysis of fundamental human needs: safety, dignity, and belonging. In ordinary life, all three are important, and healthy communities work to satisfy all three simultaneously. But in extreme situations (and existential risk is the most extreme situation conceivable), these needs can come into conflict. When safety is at stake, dignity (the freedom to make one's own choices) and belonging (inclusion in a community) must yield to the demands of safety, because neither dignity nor belonging has any meaning for people who are dead. The alignment community has made the opposite prioritization: belonging (maintaining group cohesion and identity) is treated as more important than safety (getting the truth about AGI risk right). As a result, the community's safety assessments are systematically distorted by social dynamics.

This analysis provides a diagnostic criterion. When evaluating whether a particular communication about AGI risk is truth-seeking or social-signaling, one can ask: what would this person say differently if their career, their social standing, and their group membership did not depend on the answer? If the answer is "nothing would change", the communication is probably truth-seeking. If the answer is "they would probably ex-

press much more uncertainty about alignment feasibility", the communication is probably belonging-oriented. Many alignment researchers are significantly more uncertain in private than in public: an asymmetry that strongly suggests belonging-oriented public communication.

I constructed the phrase "It is 100% possible to know that X is 100% impossible" deliberately as a diagnostic: a piece of "spicy language" designed to identify respondents who engage in social signaling over rigorous reasoning. The idea is drawn from an analogy with spicy foods in warm climates: in environments where spoilage is a risk, pungent spices help identify contaminated food. Similarly, a deliberately provocative but logically sound statement can help identify respondents whose reaction is based on social reflex rather than logical evaluation. Anyone who finds the statement "too strong" without evaluating its logical content has revealed, at no cost to me, that their standard of evaluation is social rather than logical.

> **Key Observation**
>
> The merit filter is supposed to identify people who reject reason in favor of a desired conclusion. Consider the filter applied to someone whose claim is logically grounded: the impossibility of perpetual motion follows from conservation law, which follows from symmetry, which is the foundation of all reasoning. The person applying the filter has then themselves rejected reason in favor of a desired conclusion (that AGI alignment must be possible). The filter identifies the applier, not the target.

B.2 Perfected Bias: The Allegory of the Perpetual Benefit Machine

Appendix A and Section B.1 examined obstacles to engagement with the impossibility argument at the level of the single conversation and at the level of the community. In this section I examine the deepest and most powerful obstacle: a systemic motivational structure that makes the im-

possibility of AGI alignment literally unthinkable within the paradigm that most alignment researchers inhabit. An allegory shows the mechanism: it transposes the AGI alignment situation into the context of nuclear physics in 1939.

B.2.1 Act One: The Pitch

The allegory begins with a dinner conversation. A fellow mathematician leans over and says, with great urgency, that it is going to be necessary to figure out how to make a perpetual motion machine. University researchers at Chicago are working on something called "Pile-1": some kind of free energy source. Unfortunately, the process produces serious toxins, and if used a certain way, it could make weapons powerful enough to end the world. Dangerous nations might develop it first. The only option is to be the first to build safe "Artificial General Irradiation" devices (AGI for short) and use them to prevent misuse.

In the speaker's own words, the pitch runs as follows.

Hey, um, you know, it is going to be necessary for us to figure out how to make a perpetual motion machine.

You see, these guys at the University of Chicago are working on some project with atoms. They say that they are going to somehow get unlimited energy using some machine they invented called a "reactor"; they describe it as some sort of free energy source. I believe, though they are keeping all of the details secret, that the project code-name is "Pile-1". I have no idea what it can do.

Unfortunately, some of my other friends have been researching this thing called neural understanding (no, wait, I mean understanding of *nuclear* process), and they are concerned about the real world implications.

The physicists are telling me that this thing will produce serious toxins, which are inherently hazardous to all life, via "side effects" of some sort. I forget why; they use some "R" word. Anyway, they *also* mentioned that this pollution was so toxic that if used a certain way it could be used to make weapons so powerful that they could end the world. Something serious about explosions of unusual power. They even mentioned very concerned speculation about "igniting the atmosphere", which sounds really terrible.

If the devices fell into the wrong hands, they will be used by people who are for sure much less moral and ethical than we are. We are good people. If they use it, it will likely be misused to kill everyone.

We cannot let them be the first to make these awful energy machines. We must be the ones who invent this perpetual energy source before they do, and we must somehow set it up, by our having first use of these Artificial General Irradiation (AGI) free energy devices, so that no one else can have or develop or use these terrible weapons and toxins: we *must have* the first mover's advantage. I am especially concerned that FB (that Fascist Blitz) or China (no, Japan) or some other Nazis might be the first to develop this

> sort of artificial "N-tech". We simply *cannot* let that happen! We *must* be the ones to do it first, so that maybe we can use those devices to prevent them from doing it.

The listener has been nodding along; the speaker notices, and goes straight on.

> Can I count on you to work on this project, to find some way for us to make *safe* free energy machines? We will have to be cost competitive with them, and by our controlling of our devices, we can maybe prevent all of the terrible outcome hazards. Of course, they will not even try to prevent such unethical uses of their tech; they are maybe already getting ready to advertise the benefits of their work, so as to socially justify their continued nuke capability development.
>
> So we *must* develop some sort of safety "alignment" mechanism for these machines. The FB people and the Far East nations are already involved in projects of their own; maybe they will beat out Chicago? So either we figure out a way to make safe, aligned perpetual benefit machines, right now, or we all die!
>
> Oh, and I have already talked to a bunch of my VC investor friends, told them how important this is and how much it could change the very face of the earth, how it would for sure definitely shift the balance of power for all of the national governments in the entire world, and they seemed really enthusiastic about supporting our project. They even gave me an FTX check for 100 million to start with, with a promise of more soon. We can hire whomever we like and begin our work right away. Just think of how famous we would be for liberating everyone from the dangers of nearly unlimited control and power!

Venture capital investors have provided a hundred million dollars. The work will attract unlimited prestige. The alternative is that everyone dies. You have a family to protect. What do you do?

You start working on "AGI alignment": figuring out how to make the free energy machines safe. The stakes are existential: the world depends on your success. The funding is generous. Your colleagues are brilliant. The cause is noble. You begin, feeling confident.

B.2.2 Act Two: The Plateau

Time passes. The safety work proves more difficult than expected. Complexity problems proliferate. Every approach that seemed promising encounters unexpected obstacles. Meanwhile, capability research continues to advance; the Chicago reactor is approaching "criticality". Betting pools form on timelines. A new field called "Energy X-Risk" has emerged. Some people are losing faith that the weapons can ever be made "aligned" with human wellbeing, though nobody is ready to abandon the hope that some alternative form of the technology might be safe.

The parallel to the current AI alignment situation is precise. A decade of alignment research has not produced a single method with any formal guarantee of safety. Capability research, meanwhile, advances rapidly. The alignment community has grown into a set of institutions with funding, careers, conferences, and a shared identity. Some researchers privately express doubt that the problem is solvable, but no one is prepared to say so publicly, because the implications are too threatening to the entire institutional structure.

B.2.3 Act Three: The Dinner Party

At a dinner party, the protagonist offers the optimistic view: "It *feels* to me like it should be possible to avoid egregious machine outcome misalignment, regardless of how the empirical facts shake out". Nods of approval all around. Then an unknown visitor (scruffy, uninvited, lacking the social credentials of the assembled company) speaks up:

> "Actually, it *is* 100% possible to know, today, that it is 100% impossible to build a perpetual motion machine. There is no such thing as 'free' energy or unlimited benefit. Everything has actual costs and real risks, and we cannot simply 'should' them away, no matter how much we want to".

The room descends into silence. Then pandemonium. The sequence of responses is predictable and follows a fixed pattern: "That sounds overstated"; "He sounds way too overconfident, to suggest everyone else in the world is totally wrong"; "I think we can safely ignore anyone who dares use the forbidden taboo word 'proof'". The visitor attempts to respond: "Causation and logic, as basic tools and concepts, are not infinitely powerful. They have real limits as to what they can do". But it is too late; the yelling has begun.

The protagonist tries not to notice the double bind in the third response: if no one is permitted to use the word "proof", how is anyone ever going to suggest anything, and have anyone really trust it, or make any new argument at all? The visitor's reply also says this:

> "It simply cannot be assumed that they can solve all kinds of problems. While we might desperately *want* to make and have a machine that can do at least roughly what we want, that is not always possible. Sometimes it is just inherent in the nature of the tech itself that it will create egregious outcome misalignments. I know that goes against the general feeling and opinion that 'technology is at worst neutral', and maybe somewhat more often 'good and beneficial'".

The protagonist's own misgivings emerge, and they are revealing: "But that cannot be! The tech progress is inevitable! Either we find a way to make it safe or we all die! We dare not go against the capabilities companies. They are too powerful. They will take away our funding". These misgivings are motivational, not logical. They assert that the claim must be wrong because the consequences of its being right are unacceptable; they do not engage with the claim that the technology is inherently unsafe.

The protagonist's last misgiving concerns prestige: "Nothing you are offering has even the least bit of prestige or job satisfaction attached". Nothing more is heard or remembered; everything else that occurs that evening is blotted out in an alcohol-fueled blitz of total *cognitive dissonance* (the discomfort of holding two beliefs that conflict, and the ways people avoid feeling it). No one notices the *Stockholm Effect* (the tendency of captives to come to side with their captors) happening. The motivations and incentives are simply too strong, and so reason itself will sometimes be rejected, even by rationalists who try to uphold the hype-dream.

B.2.4 The Moment of Perfected Bias

The allegory identifies a specific psychological moment at which bias becomes "perfected": that is, becomes invisible to the person who holds it. This moment occurs when the perceived desirability and necessity of a goal become conflated with its feasibility. "We must do it" becomes "We can do it" without the intervening question "Is it physically possible?" ever being asked.

> **Perfected Bias**
>
> The psychological state in which, for some goal, "we really want to do it" and "we must do it" become conflated with "we can do it" and "we could do it". The question of whether the goal is physically or mathematically possible never gets examined. It becomes unthinkable: a thought-crime. Anyone who raises it commits heresy against the shared project and is subject to social sanctions.

Perfected bias differs from ordinary bias. Ordinary bias distorts the evaluation of evidence; perfected bias prevents the question from being raised. It operates at the level of deciding which questions are permitted, not at the level of reasoning about a question. The person experiencing perfected bias is not reasoning badly about whether AGI alignment is possible; they are not reasoning about it at all. The question has been removed from the space of thinkable thoughts by the combined force of existential fear, career dependence, identity investment, and social reinforcement.

The nuclear allegory sharpens this insight with a historical detail. Before the Trinity Test, the first nuclear detonation, there was significant concern that the explosion might ignite the nitrogen in the atmosphere and destroy all life on Earth. Calculations were performed to assess this risk. The calculations were later found to contain errors; it was, in a meaningful sense, coincidence that the correct calculations produced the same reassuring answer as the incorrect ones. The Trinity risk assessment is perfected bias in action: the desire for the test to proceed was so powerful that

the risk assessment was performed under conditions that would not have been tolerated in any other context. The same dynamic operates in the AGI alignment community: the desire for alignment to be possible is so powerful that arguments for its impossibility are dismissed under standards that would not be applied to any other technical claim.

The nuclear parallel is instructive in a further way. Nuclear technology did, in fact, work: reactors produce energy, and bombs explode. The parallel to AGI lies in the *safety promise*, not in the technology's basic function. The promise of nuclear energy was "clean, safe, unlimited power". The reality was toxic waste with storage problems that remain unsolved after seventy years, weapons that threaten civilization's existence, and costs that are only marginally covered by the electricity generated over a plant's lifetime, even with heavy government subsidies. When the total costs are included (decommissioning, waste storage, the risk-adjusted cost of weapons proliferation, the long-term health effects on surrounding populations), nuclear power has been a net negative for humanity, by the accounting I make here. Accountings that count instead the deaths from accidents, and the deaths from air pollution that nuclear power has prevented, reach the opposite result (Kharecha and Hansen, 2013); the difference lies in what is counted. Yet the "clean, safe, unlimited power" narrative persisted for decades, sustained by perfected bias: the desire for the technology to be beneficial was so strong that the evidence of its costs was systematically minimized.

The AGI parallel is precise. AGI is being described as a source of unlimited benefit: the "perpetual benefit machine" that will solve poverty, cure disease, and usher in utopia. The safety promise is that alignment research will ensure the technology serves human interests. By the analysis of Chapters 1 to 4, which Volume 2 gives again formally, the safety promise is as illusory as the promise of "clean nuclear energy". The technology inherently produces effects (displacement of organic systems, accumulation of the environmental costs of its infrastructure, erosion of human autonomy) that alignment cannot prevent, because they are consequences of the technology's basic operation, not of its misalignment. In this analys-

is, the alignment researcher plays the same role as the nuclear safety engineer who promises that the waste problem will be solved: a promise made in good faith and contradicted by the physics of the situation.

The allegory also draws attention to the historical irony of the "first mover advantage" argument. In the nuclear case, the first-mover advantage narrative ("we must develop the bomb before the Nazis do") was used to justify the Manhattan Project. The Nazis, as it turned out, were nowhere close to developing a nuclear weapon. The first-mover advantage was real, but it accrued to the United States in the form of a terrifying and destabilizing weapons monopoly, not in the form of safety. The subsequent history of nuclear proliferation demonstrates that first-mover advantage does not prevent other actors from eventually developing the same technology; it merely determines who holds the initial monopoly. In the AI context, the same logic applies: being the first to develop AGI does not prevent other nations or corporations from developing it; it merely determines who bears the initial risk. Because the technology is inherently unsafe, being the first to deploy it is a liability.

The mechanism has four mutually reinforcing components. *Existential fear* creates the urgency: if we do not solve this, everyone dies. *Funding* creates the career dependence: if the problem is unsolvable, the careers and institutions built around solving it have no justification. *Prestige* creates the identity investment: the alignment researcher is "saving the world", and this identity cannot survive the recognition that there is nothing to save it from (because the threat is inherent in the technology, not contingent on misalignment). *Social reinforcement* ensures that anyone who questions the feasibility of alignment is treated as a threat to the group rather than as a contributor to the truth. Together, these forces produce a state in which the impossibility of alignment is not merely unlikely or unwelcome but literally unthinkable.

The nuclear allegory illuminates an additional dimension: the role of *race dynamics* in perfecting the bias. In the allegory, the urgency to develop safe technology is driven not only by the technology's inherent danger but by the fear that adversaries will develop it first. "We must be the ones to

do it first, so that maybe we can use those devices to prevent them from doing it". This race dynamic transforms the risk calculation: even if the technology is dangerous, failing to develop it is perceived as even more dangerous, because adversaries will develop it anyway and use it irresponsibly. The race dynamic makes the development of the technology seem *obligatory*, and any argument for halting development is reframed as an argument for unilateral disarmament in the face of a hostile competitor.

In the actual development of AI, this race dynamic is pervasive. US-China competition in AI development is routinely invoked as a reason why development cannot be slowed or stopped, even by those who acknowledge the existential risks. The argument is always the same: "If we don't build it, they will, and they will not bother with safety at all". This argument has a specific logical structure: it assumes that the only way to address the risk of a dangerous technology is to develop that technology yourself and attempt to make it safe, rather than to prevent the technology from being developed at all. Because the technology cannot be made safe, the race-dynamic argument collapses: developing the technology yourself does not make it safe; it merely ensures that the inevitable catastrophe occurs on your watch rather than your competitor's. But this is the kind of reasoning that perfected bias prevents: the conclusion is so threatening to the entire framework that the premises are never examined.

The allegory also captures the role of venture capital and financial incentives in sustaining the bias. The hundred-million-dollar grant in the story is generous, and it also creates a dependency relationship. The researcher who accepts the funding becomes, in effect, an employee of the funding source, working toward the funder's goals. The venture capitalists are enthusiastic because they recognize that the technology "will change the economy" and "shift the balance of power", and they want to be on the winning side; their enthusiasm does not come from care about safety. Their support creates the material conditions that make the research possible, but it also creates the incentive structure that makes questioning the research impossible. The researcher who concludes that the project is impossible has no funding, no career, no institutional support.

The researcher who concludes that the project is difficult but feasible has unlimited funding, unlimited prestige, and the gratitude of powerful institutions. The incentive structure is not subtle, and its effects on the direction of research are not speculative.

B.2.5 What Each Element of the Story Stands For

Each element of the allegory maps onto the present situation of AGI and AI alignment. A safe perpetual motion machine stands for aligned AGI; the Chicago reactor stands for capabilities research; radiation and the bomb stand for the toxic side effects of AGI and for its weapons; and the race against the other side stands for the race to build AGI first. Chicago Pile-1 was the world's first man-made (non-natural) nuclear reactor. The "R" word the physicists use is "radiation", and the "serious toxins" are the radioactive waste produced by nuclear power.

The slips of the tongue in the pitch let the present show through 1939. "Neural understanding", corrected to "nuclear", "Artificial General Irradiation (AGI)", and "N-tech", a conflation of "neuro-tech" and "nuclear-tech", all place the AI project inside the nuclear one. "FB", glossed in the story as "that Fascist Blitz", is Facebook; and "China (no, Japan)" names a rival of the present in place of a rival of 1939.

The pitch says "We are good people". Everyone thinks of themselves as good people, and so the factor of this being a kind of bias and self-rationalization never occurs to anyone. The hope of being first, so as to use the devices to prevent the others from building them, is another common argument among AGI capability developers: the false belief that AGI, as a hyped super-being, could somehow prevent the formation of any other AGI. It is the belief in monotheism, "one way", all over again. The other claims of the pitch (that the other side will not even try to prevent unethical uses of its technology, and is getting ready to advertise the benefits of its work so as to justify its continued development) are arguments commonly found on AI alignment forums, made by "researchers" working on these topics.

The venture investors hear the claims that the technology "will change the economy" and "will change the balance of power for all of the national governments". They know that when a major new game comes to the market, there will be whole new categories of winners and losers, and they want to be on the winning (profitable) side; by investing, they will *own* the technology, and so can capture its outputs, benefits and value. Of course they are enthusiastic about supporting the project: its researchers have voluntarily become their loyal employees and lackeys, working to their unending profit. The promise of fame for "liberating everyone" plays into the childish male geek fantasy of "saving the world", for the adoring princess. The sentence about the fantasy of saving the world is an observation of how the incentive of fame acts, not a judgment of any person's character.

The full accounting of nuclear power adds one more thing. When it is considered from a net overall benefit, cost and risk perspective, all factors included, the actual total harm to humanity and to life overall is very much greater than any partial gains (income and riches) that accrue to a very few people. A clearer case of a technology that is net more costly to civilization, but that is billed as a benefit, is hard to come by. The only reasons anyone at all tolerates "building the nuke plant" over "not building the nuke plant" are that the distribution of resources favors the decision makers, or that some government somewhere wants *fissile material* (material that can sustain a nuclear chain reaction) to make into threats to end rational civilization.

B.3 Proof Irrationality: The Epistemic Incoherence of "Anything Is Possible"

The deepest critique of the alignment community's response to the impossibility argument is epistemological, not social. It concerns an incoherence at the foundation of the community's position: a commitment to ra-

tionality held together with a refusal to accept that rational inquiry can yield certain knowledge about impossibility. This incoherence is *proof irrationality*.

B.3.1 The Crux

The crux of the matter can be stated as a series of questions. Take the belief that "anything is possible", that no knowledge can be complete or 100% certain, and that some future observation could always change everything. How is that belief not equivalent to asserting that there are no limits to what can be done with causation, algorithms, reason, logic, and mathematics? How does the person who holds this view account for the already known and proven limits: the Halting Problem, Rice's Theorem, microstate amplification in nonlinear systems, the Heisenberg Uncertainty Principle? How is the assertion that "nothing is perfectly knowable" not a rejection of the notion of symmetry, and therefore of reason and physical lawfulness themselves?

These questions are asked to be answered: they point to a logical problem at the foundation of the alignment community's epistemology.

One more question belongs with them. Is there any way to assert that the notion of reason and rationality is coherent without also implying a completed acceptance of symmetry as an absolute, and so also accepting as absolute the notion of causation, some actual real limits to what can be done with causation alone, and a notion of physical lawfulness? The question of whether reason can be coherent without symmetry being accepted as absolute is a core epistemic crux. No point of proof, and no notion of rigor or reason, can obtain beyond the crux unless the crux is resolved first. As near as I can tell, every notion of reason, logic and knowability, up to the point of completed certainty, depends on whether the notions of proof and symmetry are fully bijective in their truth: matched one to one, so that each fully determines the other. A belief in *Bayesianism* (the view that knowledge is a matter of degrees of belief, updated by evidence according to the rules of probability) does not even begin to cover this, for

even the notion of *object permanence* (the understanding that things continue to exist when not being perceived), and of the consistency of the perceived world, cannot be assumed without at least some notion of symmetry as a completed, 100% accepted prior.

B.3.2 Symmetry and the Basis of Reason

Section B.1.1 established that symmetry is the common foundation of both physical law and logical truth. Every valid inference preserves truth; truth preservation under transformation is a form of symmetry; the consistency of any mathematical system depends on this symmetry. Physical laws are expressions of symmetry (via Noether's theorem). The concept of causation itself is a symmetry: a persistent pattern of observation where, under the same conditions, the same outcomes obtain.

Put most simply, symmetry is a sameness of content where there is a difference of context. Causation, kept maximally simple, is a persistent and consistent pattern of observation in which everyone notices that first Q happens, and then P happens; the term refers to the implied lasting symmetry of perceptual patterns (content) over many people and situations (context). This reaches even to perception itself. If there is zero symmetry between the inputs and the outputs of a communication channel (between what a sensor is sensing and the signal it transforms that into), then the very notion of sense is obviated. There can be no truth coherence in mathematics (let alone proof), and no science of experiment based on perception and observation, if even the regularity of "first we see this" and "then we see that" (causation) is rejected. The notion of pattern and regularity is itself an unquestioned assumption of symmetry, as a basis for reason and rigor of thinking; and reason is itself a kind of abstraction of perception, which also tacitly assumes symmetry.

Every form of knowing rests on symmetry; to reject it, or even to treat its application as uncertain or provisional, undermines the basis of all knowing. One cannot coherently claim to be a "rationalist" (one who upholds reason as the basis of knowledge) while simultaneously claiming

that the foundation of reason (symmetry, and the certainty it provides) is not to be fully trusted. Yet many in the alignment community implicitly adopt this position when they assert that "nothing is 100% knowable" and "everything might turn out to be possible".

The self-referential difficulty is stark. Is it 100% knowable that nothing else is 100% knowable? If the answer is yes, then the claim is self-defeating: at least one thing (the claim itself) is 100% knowable, which contradicts the universal assertion. If the answer is no, then the claim is not itself certain, and we cannot rely on it as a basis for dismissing impossibility arguments. Either way, the position is logically untenable.

Coherence holds in the same way. Every concrete argument presumes at some point that objectivity is meaningful, and many people in the AI community will go as meta as they can in order to be able to dispute anything. So at some point it has to be said plainly: one can call the notion of meaningfulness coherent or incoherent, but whichever way one describes it, it is coherent. The same holds for the coherence of the world. Divide subjectivity from objectivity. One can go fully *idealist*, holding that there is no objective world, only the subjective; then the question is whether the notion of coherence applies to the subjective. Or one can go the other way, holding that there is only the objective, only the real, and treating subjectivity as an *epiphenomenon* (a by-product of a process, with no effect back on that process), which is the *materialist* way of thinking; then the question is whether the world is incoherent. For one who holds that it is, mathematics, science and any coherence of physical law go out the window, and that person's mental model is so incoherent that no conversation is possible, because there is no knowing whether that person uses the word "dog" in the same way from one moment to the next. The view that we only build coherent representations of an incoherently complex world, filtered through our biases, follows the direction of Kant, and much of my work is a refutation of Kant.

B.3.3 The Self-Defeating Exception

The practical consequence of this epistemological incoherence is a selective application of certainty. The alignment community is willing to accept 100% certainty when it favors the possibility of its project. The laws of physics are reliable enough to build the computers that run AI systems. The principles of mathematics are reliable enough to prove the theorems on which machine learning is based. The methods of engineering are reliable enough to construct the infrastructure that powers AI research. But the same community refuses to accept 100% certainty when it threatens the possibility of its project. The same laws of physics, the same mathematical principles, and the same engineering constraints that make AI systems possible also impose limits on what those systems can do and on what can be guaranteed about their behavior.

This selective acceptance of certainty is motivated reasoning, not a principled epistemological position. Certainty is accepted when it is useful and rejected when it is threatening. The community maintains its self-identification as "rationalist" by never making this selectivity explicit. It treats the acceptance of certainty (in physics, mathematics, engineering) as so obvious that it does not need to be stated. It treats the rejection of certainty (regarding impossibility results) as a principled commitment to epistemic humility.

The result is that the alignment community, for all its commitment to rationality, is not behaving rationally with respect to this specific question. It is behaving in a way that is predictable from the perspective of motivated reasoning, identity-protective cognition, and perfected bias, but not from the perspective of logical consistency. That the alignment community is not behaving rationally with respect to the possibility of AGI alignment is the central epistemic fact about the alignment community's relationship to the impossibility argument.

The selectivity can be stated exactly. The consistent claim is that "you cannot 100% know today that we may not someday do X". Eventually, X will happen to be some notion of "align AGI to human interests"; it seems so simple, and so desirable, to believe that we can eventually do what we want. X will never be anything like "prove to 100% confidence that AGI alignment is actually completely structurally impossible, for any real notions of 'aligned' and 'safe'". Allowing X to take only good claims, never bad ones, is exactly and logically equivalent to *sometimes* rejecting symmetry as an applicable notion in reason, logic and knowability, and there is no justification for this "sometimes" other than the illogical bias of motivated reasoning. In this way the rationalists become irrational, at least where the AI hype is concerned, and the AI alignment research community is already behaving in a way that is fully, 100% knowably and verifiably, illogical and irrational. The alignment community is already behaving illogically well before any actual reasoning about how to show and know that AGI is inherently unsafe is even described. People will quibble about definitions of words all day while the planet burns around them. The demand that the argument be put in "formal language" is a nonsense issue: everyone already knows from basic computer science that algorithms can be expressed in any language, and the same holds for proofs. If reason and logic cannot be used to establish a proof of the impossibility of something as basic as a perpetual energy machine, then handling any reasoning about something more abstract, such as a perpetual benefit machine (which AGI utopia is currently being billed as), is not possible for this group of people. For the rationalist community to keep its title as rational, most of its members will need to rise to the standard of actually being rational, which also means rejecting the false notion that "nothing is 100% certain" and "everything is possible".

B.3.4 AGI as Perpetual Benefit Machine

Proof irrationality takes a more specific form when applied to the concept of AGI itself. The question can be put directly: in what important sense is AGI *not* being described as a "perpetual benefit machine", a machine that

can do anything that is doable, make anything that is makable, and solve any problem that is physically solvable? How does AGI not eventually become equivalent to an assumption of limitless intelligence?

The alignment community's own discourse provides abundant evidence that AGI is indeed conceived in these terms. The standard framing is: "Everything will probably be better once we have AGI, if we can align it first". The utopian scenario is a machine that solves all humanity's problems: poverty, disease, environmental degradation, death itself. The dystopian scenario is the same machine, unaligned, that destroys everything. In both scenarios, the machine is assumed to be effectively omnipotent within the domain of physically possible outcomes. The only question is whether its omnipotence serves human interests or subverts them.

This framing has a specific logical consequence. Once AGI is conceived as a perpetual benefit machine (a machine whose benefits are unlimited and permanent), the question of whether it can be "aligned" is structurally identical to the question of whether a perpetual motion machine can be built. Both involve the assumption that a single technology can produce unlimited benefits indefinitely, without corresponding costs, limits, or degradation. The impossibility argument of Chapter 1 establishes, and Volume 2 proves again formally, that this assumption fails for AGI for the same structural reason it fails for perpetual motion. In both cases the assumption requires violations of conservation principles: the conservation of energy in the physical case, and the conservation of information and the limits of causation in the computational case.

A second question sharpens the point. Is there any proposed method of AGI alignment that does not depend, directly or indirectly, on causation and modeling? That is: is there any alignment method that does not operate through some form of algorithm (a causative model of a system that implements alignment error correction), or through some form of economic or institutional incentive (a causal mechanism for shaping behavior)? If all proposed alignment methods depend on causation, then the known limits of causation (the Halting Problem, Rice's Theorem, the data processing inequality, the uncomputability of general behavioral proper-

ties) apply to all proposed alignment methods. No amount of engineering cleverness can circumvent a mathematical limit, any more than engineering cleverness can circumvent conservation of energy.

It does not seem reasonable to assume that an AGI will somehow "magically know" what our best intentions are, so as to do the right thing for our benefit. For a machine to "know", it must have some process, and since "machine" implies repeatable process, the notions of symmetry and causation both apply. Even the notion of algorithm, and even logic itself, imply some equivalence to causation in this sense. Anything done with machines, and therefore with engineering, directly or indirectly, is always about causation (and not, for example, magic). These questions will be either debated or ignored. Ignoring them amounts to a tacit acceptance of what they imply: that the notions of causation and of symmetry apply to machines, whatever their type (neural networks included, however modeled).

The definition of "causation" in this context is deliberately minimal: it is merely a persistent, consistent pattern of observation in which, under the same conditions, the same outcomes obtain. The minimal definition of causation is the definition of symmetry applied to temporal sequences. It is the foundation on which all of engineering, all of science, and all of prediction rest. Any proposed alignment method that claims to operate through some mechanism other than causation would be claiming to operate through magic: through a process that does not exhibit regular patterns and therefore cannot be engineered, tested, or relied upon. No serious alignment proposal makes such a claim. Therefore, all serious alignment proposals are subject to the known limits of causation, and the impossibility argument, which establishes that those limits preclude reliable alignment, applies to all of them.

The same ground supports a proof of impossibility. Every form of logic, of reasoning, and of reasonable argument depends, directly or indirectly, on symmetry; no valid method of reasoning or proof can violate symmetry with impunity. At least some arguments, proofs and results depend only and exactly on symmetry, and these make something actually

"100% knowable". To reject the idea that anything at all is 100% knowable is therefore, at least implicitly, and also actually, to reject the idea of symmetry, and with it the idea that any form of reliable, trustworthy reasoning is possible. Where exactly nothing is fully knowable, not even symmetry is an allowed assumption; and where every form of knowing depends on symmetry, there is no way to know, as a special case, that nothing is knowable. Valid reasoning and violation of symmetry are mutually exclusive. So any statement P that inherently contradicts symmetry (directly, through false syllogism, or indirectly, through incorrect and non-relevant definitions) yields a fully knowable truth: that P is 100% known to be fully and actually impossible, in the strict, correct and absolute sense of the terms.

In the same way that all forms of perpetual motion eventually require an assumption that violates symmetry, in the form of the law of conservation of matter and energy, all notions of AGI alignment eventually require the assumption of something that causation simply cannot do. Causation is a symmetry concept. To require from causation something non-causative, outside the scope of what causation allows, is again to violate symmetry: it requires some place where exactly the same initial conditions sometimes give different output conditions. The notion of predictability itself is then compromised, and with it the ability to make even predictions like "AGI will always do what is in the human best interest". By this contradiction of the symmetry it requires, it is "100% knowable" that AGI alignment is impossible, in the same basic way, and to exactly the same standard, as it is known that perpetual motion is 100% impossible. It all comes to rest on what causation cannot do: the limits of modeling, of what can be done with game theory and with imperfect information, and of algorithms and computability. Aside from symmetry and its violation (contradiction), nothing else is available as the valid and accepted distinction between truth and falsity. All of the reasoning about symmetry and causation, together with all of the detailed side arguments that are components of any actual reasoning about the contradictions inherent in AGI safety, is still too much for most people to hold together at once.

B.3.5 The Content-Context Problem

A final dimension of the epistemological critique concerns the relationship between content and context in intellectual communication. The alignment community's frequent advice to "let the content do the talking" (to present the argument in its purest logical form and trust that its merits will be recognized) reflects a naïve view of how ideas are actually received in social settings.

Content never occurs without context. The social context in which an argument is presented (who presents it, where it appears, what precedes and follows it, who endorses or dismisses it) shapes its reception at least as much as its logical content does. An advertisement for baby food has a different meaning next to an article about food safety than it does next to an article about corporate fraud. Similarly, an impossibility argument has a different reception when presented by an established professor at a prestigious institution than when presented by an independent researcher on a public forum. The belief that content can be evaluated independently of context is itself a form of naïveté, and it serves the interests of those who control the context, because it allows them to shape reception while maintaining the appearance of evaluating arguments purely on their merits.

Saying that content never occurs without context is a recognition that the strategy for communicating the impossibility argument must account for social realities as well as logical ones, and not a counsel of despair. Formal proof, however airtight, is a necessary but not sufficient condition for acceptance. The sufficient conditions include social credibility, institutional support, and a context in which the argument can be encountered without the interference of the defensive mechanisms described in these appendices. Creating those conditions is a social and strategic challenge as well as a logical one.

This paragraph is for the reader who will present this argument to a technical community. The strategy follows from what has been said. The usual epistemic move is to start with definitions, but adverse listening will

try to pick apart every aspect of every definition, so the definitions cannot simply be presented in plain text. Coherence must first be reached around even more basic epistemic technique, so as to separate the social and political aspects from the rational ones cleanly: to separate, implicitly and consistently, the participants who treat the exchange as conflict from the collaborators and communicators who treat it as the correction of mistakes (conflict theory and mistake theory, Section A.1.3). The better course is to ask the kinds of questions that let the community obtain the needed definitions, and methods of reasoning, on its own: questions that converge on closed forms. The definitions it reaches in that way are ones it assents to, because it has worked out their reason and reasonableness itself; no one can do the understanding and comprehension for another. For the result to become unarguable, it must be based on the community's own assertions. That in turn requires keeping at least some control of the context: not holding all of the debate on the community's own forum and on its terms, not replying to every remark, moving directly to inquiry, and, wherever possible, holding the conversation one on one.

B.3.6 Logic and Relevance

Establishing the formal contradiction of long-term AGI safety needs both logical parts and relevance parts. An argument built to focus on logical validity alone ends as an overly simplified toy model, and then everyone discounts it as "not relevant", because the toy model does not correspond to the real world; in this case, mathematics without relevance is useless. Relevance has its own discipline. People tend to think that they can define terms in whatever way they like, as if it did not matter; yet just as much discipline is needed in how terms are defined, so that the right relevance is available, and that is not a common skill.

There is therefore a balance between how an argument is set up epistemically and what it means to have a context of truth. More time spent scaffolding the needed skills makes the argument more airtight, and very much longer overall. Abbreviating those skills, and the reasoning behind both the argument and its methods of reasoning, wastes a great deal of

bandwidth on needless confusions and mistakes (or on politics, bias and motivated reasoning). The only low-bandwidth option is a summary in outline, which is not a proof, and which also displeases everyone as unsatisfying. The overall inferential distance cannot be made shorter without short cuts based on prior knowledge, which most participants in the alignment forums do not have, or choose to ignore. A proof of anything can be fielded only where the listeners are reasonable, and are willing to commit to the notion of rigor; there are no half measures here, and no trust of any kind otherwise, particularly where actual responsibility for safety is concerned. Hype will not cut it here.

Any argument of the form "I do not believe you could convince other people of this" is a *social argument*: it rests on an impression of what other people's impressions will be. What other people will think has zero relevance to an argument about the relationship between AI and the physical universe, AI and matter. Arguing from what others will agree to is like ants debating what other ants will think while the boot is coming down. What they think does not matter; the boot is coming down, and its physical reality is the whole deal. The politics of ants do not matter. The argument is entangled in social questions only in that it has to be brought forward; for the basis of the argument, social tractability (how easily an argument can be brought to be accepted) and normative statements about what anyone should do are irrelevant to the facts. A state legislature once moved to declare the value of pi by statute; mathematics cannot be legislated. Whether an argument is socially tractable is a concern of presentation, not of proof, and what is at issue here is a proof: an observation about the relationship between information theory, chemistry and evolutionary biology.

Appendix C

Rhetoric, Clarity Politics, and Fallacy

The rhetorical tactics and argument forms that deflect the impossibility claim

C.1 Clarity Politics: The Rhetorical Battlefield

Appendices A and B examined cognitive and motivational obstacles to engaging with the impossibility argument. This appendix examines the *rhetorical* tactics by which the AI safety and capabilities communities deflect uncomfortable arguments without engaging their substance. The name for the whole set of these tactics is *clarity politics*. Clarity politics is the tactical use of rhetoric (the redefinition of key terms, the shifting of the burden of proof, and the concreteness dilemma, each described below) to neutralize an argument in conversation without evaluating it. The analysis is sharp and is intended to be: the stakes are existential, and rhetorical misdirection in the context of existential risk is dangerous.

C.1.1 The Complicity of Safety Research

The social critique applies to AGI capabilities developers and also to the alignment research community itself. The critique of capability developers is relatively straightforward. They are aware of the risks, and they deploy spin and "align-washing" (claims of alignment made for public relations, on the model of greenwashing) to divert attention from those risks. They continue development because of private profit incentives and personal bias toward replacing organic processes with artificial ones. These critiques are widely shared even within the alignment community.

The more uncomfortable critique is directed at the alignment researchers themselves. My claim is that safety research, however well-intentioned, enables capability development. By sustaining the belief that AGI

can in principle be made safe, alignment researchers provide the intellectual cover that allows capability developers to continue their work. The alignment researcher's implicit message to the public is: "This technology is dangerous, but we are working on making it safe", which translates, in practice, to: "Do not worry; do not demand that development be stopped; experts are handling it". No method of alignment can provide reliable safety guarantees; this message is therefore false, and the alignment community is complicit in the very risk it claims to be mitigating.

This position is uncomfortable to articulate, and it is acknowledged as such. Many alignment researchers are motivated by concern for human welfare. The claim is that they are caught in a structural trap, not that they are insincere. Their sincere desire to make AGI safe leads them to produce work that functions, regardless of its content, as a signal that safety is achievable, and that signal enables the continuation of capability development. The sincerity of the motivation does not change the function of the output.

C.1.2 Identity-Based Resistance

An observation often attributed to Upton Sinclair holds that it is difficult to get someone to understand something when their salary depends on their not understanding it. In this context the observation extends to a stronger claim: it is effectively impossible to get someone to understand something when their *identity* depends on their not understanding it. The alignment researcher's identity is built on the premise that alignment is a meaningful research problem: hard, perhaps, but solvable in principle. To accept the impossibility argument is to accept that one's career, one's expertise, and one's self-conception as someone who is "working on the most important problem in the world" are all based on a false premise. Such acceptance is more than an intellectual adjustment: it is an identity crisis.

The psychological literature on identity-protective cognition (Kahan, 2013) and motivated reasoning (Kunda, 1990) provides substantial support for this general pattern. People evaluate evidence for its implications for their identity and group membership as well as for its logical merit. When the evidence threatens identity, it is more than discounted: it is actively reinterpreted to support the pre-existing commitment. This reinterpretation can be entirely unconscious: the person believes they are evaluating the evidence on its merits, while in fact their evaluation is being shaped by identity-protective cognition at a level below conscious awareness.

The practical implication is sobering: better argumentation will not overcome identity-based resistance. The obstacle is that the argument threatens something more fundamental than a belief: it threatens a self-conception. The obstacle is not that the argument has been presented without enough clarity. No amount of clarity will make a person voluntarily undergo an identity crisis, and asking them to do so is not a reasonable expectation. The conclusion is that the argument's audience should be chosen with care, not that the argument should be abandoned: those whose identity is not bound up with alignment research are more likely to evaluate the argument on its merits.

The identity-based resistance is further strengthened by the community's investment in what might be called the "hero narrative". The alignment researcher conceives of themselves as working on "the most important problem in the world", a narrative that combines existential urgency with personal significance. The geek fantasy of "saving the world", which has deep roots in science fiction and technology culture, provides a powerful identity template. The researcher is the protagonist of a story in which their technical skills are what is needed to avert catastrophe. This narrative is psychologically intoxicating, and it is reinforced by the community's culture, its funding structures, and its public messaging. To accept the impossibility argument is to accept that one is not, in fact, the hero of this story: that the story itself is a fiction, and that the real situation is one in which heroism of this kind is not available.

The pattern extends beyond individual psychology to institutional dynamics. Organizations that have been built around the premise that alignment is a solvable problem (research institutes, grant-making foundations, academic departments) have institutional interests in the continued perception that the problem is solvable. Their funding, their staff's employment, and their organizational identity all depend on that perception. Even if every individual within such an organization privately came to believe that alignment is impossible, the organizational incentives would resist the public acknowledgment of this conclusion, because that acknowledgment would mean the end of the organization. Institutions, like individuals, have self-preservation instincts, and those instincts operate independently of the truth.

The pattern requires only that people respond to incentives in the ways that people normally respond to incentives, which is to say, in ways that are heavily influenced by self-interest, identity, and social belonging, even when those influences are entirely unconscious. It is not a conspiracy theory, and it does not require anyone to be acting in bad faith. The alignment community is composed of intelligent, well-meaning people who are concerned about the future of humanity. But concern, however sincere, does not immunize against motivated reasoning, and the motivational structure of the alignment community is designed, structurally rather than intentionally, to prevent the one conclusion that would require the community to disband.

C.1.3 Terminology Manipulation

A specific set of rhetorical tactics has been observed in the alignment community's response to the impossibility argument. The most consequential is the systematic redefinition of key terms.

The meaning of the term "existential risk" has shifted. In its original and common-sense meaning, it refers to the risk that human beings may cease to exist. In some alignment-adjacent discourse, it has been redefined to mean the risk that *technological civilization* may cease to exist, a defini-

tion that allows for the continuation of artificial systems even if organic life is displaced. This redefinition is not innocent: it converts a risk to humanity into a risk to a technological infrastructure. It thereby aligns the concept with the interests of technology developers rather than with the interests of the humans the concept was originally meant to protect.

Similarly, the scope of the term "aligned" has been manipulated. In ordinary usage, it means "aligned with the wellbeing of all of humanity" or "aligned with all of life". In practice, it often means "aligned with the preferences of the system's operators" or "legally safe for the deploying corporation". The term "safe" undergoes the same transformation: from "safe for people" to "safe from liability". Each redefinition does the same thing: it substitutes a weaker, more achievable goal for the stronger goal that the public discourse is ostensibly about, while using the same terminology to maintain the appearance that the stronger goal is being pursued.

C.1.4 The Concreteness Dilemma

A particularly effective rhetorical trap operates through what might be called the concreteness dilemma. If the impossibility argument is presented in specific, concrete terms, with detailed technical analysis of particular alignment proposals, the response is: "Your analysis is too narrow. You have shown that *this particular approach* fails, but that does not show that *all* approaches fail. The scope and generality of your argument is too limited to matter". Nitpicking and definitional disagreements then offer endless opportunities for distraction. If, on the other hand, the argument is presented in general terms, establishing structural reasons why no alignment method can succeed, the response is: "Your argument is too vague. You have not engaged with the specific details of any actual alignment proposal. You have not actually said anything concrete, and therefore your argument is not relevant".

The dilemma is a classic double bind: heads I win, tails you lose. As the one who makes the impossibility claim, I cannot win, because the goalposts are moved depending on the level of specificity chosen. The di-

lemma works to ensure that the argument is never evaluated, not to evaluate it: whatever form the argument takes, a reason can be found to dismiss it without engaging its substance. The dilemma is the rhetorical equivalent of a merit filter that rejects all inputs regardless of their content.

The limited bandwidth of the forums where these discussions typically occur makes the dilemma worse. Modern social media and forum-based discourse is optimized for brevity and rapid exchange, not for sustained, careful argument. The inferential distance required to communicate the impossibility argument (the steps of background a listener lacks) is large, the conceptual volume is substantial, and the precision required in definitions is high. None of these features is compatible with the expectation that an argument can be understood on a single pass through a forum post. Important formal work of historical significance (Gödel's incompleteness theorems, which required the construction of an entirely new self-referential language, or Bell's theorem, which required patience with an unfamiliar formal framework) would not survive the "sniff test" currently applied to new claims on alignment forums. The expectation that everything should be immediately comprehensible, stated briefly, and compatible with existing frameworks is a structural bias against novel results.

C.1.5 Domain Switching and the Technological Omnipotence Assumption

The rhetorical dynamics observed in the alignment community are instances of a broader pattern in which the technology industry's habitual assumptions shape what people perceive to be possible. The most consequential of these assumptions is what might be called the "technological omnipotence assumption": the belief that if technology can accomplish remarkable feats in one domain, it can accomplish remarkable feats in any domain, given sufficient effort and resources. This assumption is so deeply embedded in Silicon Valley culture that people typically hold it without ever putting it into words.

The writer Scott Alexander captured the underlying fallacy by imagining how media might cover other domains using the same domain-switching logic (Alexander, 2022): in his example, a climate scientist who cannot predict a shooting in his own neighborhood is treated as unreliable on climate predictions. The absurdity is obvious when applied to climate science, but the same structure goes unchallenged in AI discourse. The claim that "we should, in principle, be able to get a machine we build to do at least roughly what we want" sounds reasonable because it generalizes from everyday engineering experience. But it smuggles in an assumption that has no basis in the formal analysis: that the class of things machines can be engineered to do includes the specific property of being reliably aligned with human interests over extended time horizons. The generalization from "machines can perform specific tasks" to "machines can be permanently safe" crosses a formal boundary that the impossibility argument identifies and characterizes.

The rhetorical pattern also reflects the structural relationship between capability developers and safety researchers. That relationship parallels the one between regulated industries and their regulators in other sectors. There is a revolving door of personnel and a shared culture and set of assumptions, and the "regulators" (safety researchers) depend on the "regulated" (capability developers) for funding, data access, and career opportunities. In such an arrangement, the structure compromises the regulatory function, regardless of the individual integrity of the people involved. The alignment researcher who wants to make AGI safe is nonetheless embedded in an institutional structure that requires them to believe safety is achievable, because if it is not, the entire structure has no justification, and the funding disappears.

C.1.6 The Inversion of the Burden of Proof

The burden of proof has been manipulated in the same way as the key terms of Section C.1.3. In any other domain involving public risk (pharmaceuticals, nuclear energy, chemical manufacturing, aviation), the bur-

den of proof falls on the developer to demonstrate safety before deployment, not on those who question it to demonstrate danger after the fact. In pharmaceuticals, the company must prove the drug is safe through controlled trials. In aviation, the manufacturer must certify airworthiness through rigorous testing. In nuclear energy, the operator must demonstrate safety compliance through independent review. The principle is the same in every case: the party introducing the risk bears the burden of demonstrating that the risk is acceptable.

In the AI alignment community, this standard has been inverted: those who question AGI safety are expected to provide formal proof of impossibility, while proponents of alignment are not required to provide formal proof of possibility. That proof of impossibility is held to a mathematical standard of certainty far higher than the standard to which safety claims are held. Meanwhile, the developers and alignment researchers are permitted to proceed on the basis of aspirational research programs that have produced no formal safety guarantees whatsoever. The alignment community has participated in this inversion by accepting the framing in which "more research is needed" is treated as a justification for continued development rather than as a reason to halt development until the research is complete.

The asymmetry is compounded by the equivocation on "impossible": those who question it are asked to meet a mathematical standard of certainty, while proponents are permitted to operate on the basis of intuition, analogy, and hope. When they use the term, they are held to a mathematical standard: they must demonstrate logical impossibility in the strongest possible sense, accounting for every conceivable alternative, with no room for doubt. When proponents use terms like "possible" or "feasible", they are permitted a much weaker standard: vague intuitions about future research directions, analogies to other engineering successes, and expressions of hope are treated as adequate justification for continuing to build systems that pose existential risk. The burden of proof is unequal, and the inequality runs one way: it structurally favors continued development regardless of the evidence.

The phrase "the case for the risks has not been clearly laid out" is a perennial favorite among those who wish to avoid engaging with the impossibility argument. Decoded, the phrase means: "The argument has not been laid out to the satisfaction of my biased judgment, my willingness to pay attention, and in a way that is amenable to my skill in debate and rhetoric". The call for "clarity" is a tactical move designed to force the argument into a format where it can be more easily picked apart, distracted from, or buried in definitional disputes; it is not a genuine request for understanding. The demand for clarity is, paradoxically, a tool of obfuscation.

C.2 The Marketing Problem

Existential risk is difficult to communicate in a culture saturated by marketing; this problem is distinct from clarity politics but related to it. Standard approaches to building public trust and engagement (telling personal stories, demonstrating emotional range, showing happy outcomes, connecting with community, issuing invitations to explore) are designed for commercial contexts where the goal is to associate a brand with positive feelings. They are structurally mismatched with the communication of existential risk, where the message is "here is something terrible you cannot avoid", not "here is something good you should want".

The mismatch shows at every level. Marketing says: "Show happy outcomes". But in existential risk communication, there are no happy outcomes to show; the message is that the current trajectory leads to catastrophe. Marketing says: "No current drama". But the situation *is* drama, an unfolding crisis. Marketing says: "Demonstrate your character and competence". But existential risk is not about the messenger; making it about the messenger's character is a distraction from the content. Marketing says: "Connect to people; show that you care". But the care in question is for the biosphere, for the long-term future of life, for things that cannot be reduced to interpersonal warmth; it is not care for people's feelings.

More troublingly, the very techniques of marketing (trust hacking, emotional engagement, community building, calls to action) are the same techniques that sustain the delusion that the impossibility argument is meant to dispel. The alignment community is, in effect, a marketing operation: it sells the idea that AGI can be made safe, using the standard toolkit of emotional engagement, community identity, and prestige signaling. Any attempt to counter this message using the *same* toolkit risks becoming just another flavor of the same marketing, differing only in its conclusions. The impossibility argument cannot be marketed in the same way that alignment research is marketed, because marketing works by generating positive associations with a product, and the impossibility argument has no product to sell. It offers no career, no funding, no prestige, no hopeful vision of the future; it offers only the uncomfortable recognition that a widely desired goal is unachievable.

The mismatch is a strategic problem, and more than a personal frustration. It means that the impossibility argument is structurally disadvantaged in the competition for public attention, regardless of its logical merits. Truth, in a marketplace of ideas shaped by marketing dynamics, has no inherent advantage over well-marketed falsehood.

The problem goes deeper than the mismatch between marketing techniques and existential risk communication. The marketing advice answered above comes from a standard marketing framework, the "6 Message Strategy" of the Lighthouse Alliance, and that framework has six categories of messaging. They are stories (demonstrating emotional range and positive outcomes), lifestyle (connecting personally with the audience), questions (creating engagement through directed inquiry), philosophy (sharing values and guiding principles), results (showing client transformation from "pain before" to "paradise after"), and invitations (offering opportunities to explore further). Each of these categories is designed to build trust in a person and a brand, leading ultimately to a commercial transaction.

Every one of these categories is structurally mismatched with existential risk communication. "Show your real life" makes the message about the messenger, which is both irrelevant and dangerous: the messenger becomes a target rather than a conduit for the message. "Demonstrate positive outcomes" is impossible when the message is that the current trajectory leads to catastrophe, and the only "positive outcome" is to avoid that catastrophe by abandoning a project that many people have invested their careers in. "Create engagement through questions" risks being experienced as manipulation when the questions lead to uncomfortable conclusions. "Show your values" in this context means expressing care for the biosphere and for organic life over technological infrastructure, which places me, and anyone who presents this argument, outside the value system of the technology-oriented audience. "Show results" in the "pain to paradise" format is impossible because the impossibility argument does not offer a paradise; it offers the recognition that a particular route to paradise is a dead end, which is experienced as loss, not as liberation.

The deeper issue is that marketing, as a way of communicating, is at root a way of generating positive associations. It creates feelings of trust, excitement, belonging, and aspiration, and then attaches those feelings to a product or a project. The impossibility argument generates none of these feelings. It generates discomfort, threat, loss, and the recognition of severe constraint. No marketing framework makes constraint feel aspirational. The alignment community, by contrast, *is* marketable: it offers hope, belonging, prestige, and the exhilarating feeling of working on "the most important problem in the world". In a competition for attention between a message that offers hope and a message that offers constraint, the hope message wins by default, regardless of which message is true.

Communicating from outside the dominant culture of the audience is a further, specific challenge. For a presenter whose ways of communicating differ from the audience's, as mine do, the standard advice to "connect with people", "demonstrate emotional range", and "show your sense of humor" is potentially counterproductive. Making light of planetary-scale life termination does not seem respectful. Connecting personally with an

audience requires social bandwidth that I, like any such presenter, do not have. Different audiences require different communicative approaches, and any choice of approach will be condemned by someone as inauthentic. The limits on how I can communicate with the audience are constraints on the communication space, not excuses. The constraints limit the available strategies, and the constraints help explain why the impossibility argument has not been communicated through the channels that the alignment community finds most natural and persuasive.

C.3 Rhetorical Fallacies That Protect the Program

The preceding sections have examined social, psychological, and philosophical dynamics. In this section I catalog specific rhetorical patterns: fallacious argument forms that recur in engagements between impossibility proponents and the alignment community. These patterns function as cognitive antibodies: they provide ready-made responses that allow the community to dismiss impossibility arguments without engaging with their substance.

C.3.1 The "Cursory Look" Dismissal Pattern

Perhaps the most common response to impossibility arguments is the cursory dismissal: the listener inspects the argument briefly, picks out a surface-level feature that "sets off alarm bells", and moves at once to counter-assertion without engaging the substance. The pattern is: read quickly, feel discomfort, state that the conclusion seems implausible, and then reassert the prior belief that alignment is possible.

This pattern differs from legitimate skepticism in that it makes no attempt to identify where, specifically, the argument fails. The response is to the *conclusion* (which is unwelcome) rather than to the *reasoning* (which is not examined). A typical engagement of this form will produce multiple unsupported assertions. The engagement asserts that there is a "broad basin of attraction for acceptable outcomes", that an aligned AGI would

"set up our environment for humanity's long-term flourishing and then shut itself off", and that current research programs are "making progress". None of these assertions is accompanied by supporting argument, evidence, or engagement with the specific impossibility claims being advanced.

The "cursory look" pattern is self-reinforcing: because the engagement is superficial, it never reaches the depth at which the impossibility arguments become compelling, which in turn confirms the initial impression that they are not compelling. The pattern can persist indefinitely, with each cursory engagement providing further apparent evidence that the arguments are unpersuasive.

C.3.2 The Basin of Attraction Fallacy

A specific and recurrent form of question-begging in alignment discourse is the appeal to a "basin of attraction" for safe outcomes. In the study of dynamical systems, a basin of attraction is a region of states that a system, once inside, keeps returning to. The argument takes the form: there exists a broad basin of acceptable outcomes; if we align an AGI to a goal within this basin, it will be resistant to leaving it; therefore alignment is tractable.

Every element of this argument assumes what it needs to prove. The existence of a non-empty basin of safe outcomes is what the impossibility argument denies. Because there are no long-term safe outcomes of AGI existence, the "basin" is of zero extent and contains nothing. The conditional ("if we align an AGI to a goal within this basin") is vacuously true but operationally meaningless. The appeal to the basin restates the conclusion in spatial metaphor.

The further proposal that a well-aligned AGI might "set up our environment for humanity's long-term flourishing and then shut itself off" adds an assumption of perfected altruism. The assumption is that an artificial agent of superhuman capability will voluntarily sacrifice its own existence for the benefit of beings with which it shares no substrate. The assumption of perfected altruism is the assumption of an Artificial God: an

entity that is omnipotent, omnibenevolent, and willing to self-annihilate. It is a theology, not a research program. And like other theologies, it is unfalsifiable in principle, since any failure can be attributed to insufficient faith (or, in the secular version, insufficient safety research).

C.3.3 The No-Proof Fallacy

A particularly insidious rhetorical pattern, which I term the *No-Proof Fallacy*, operates by attacking examples in order to dismiss general claims. It has two variants.

In the first variant, a proponent posits a general class of harmful outcomes (the harm class, X) and, to make the argument concrete, provides a specific example (the example, R). An antagonist then argues that the example R is unlikely, for some specific reason (the reason given, Q), and concludes that the harm class X is therefore unlikely. The logical gap is enormous: the refutation of one example does not refute the general class. But the pattern is socially effective because it gives the appearance of having engaged with the argument.

In the second variant, a proponent *proves* that a desirable outcome (the goal, Z) is impossible via any means, and provides a specific indication of this impossibility (the indication, S). The antagonist argues that the indication S might be circumvented by some unspecified method (the workaround, M), asserted to exist in principle but never demonstrated, and concludes that the impossibility of the goal Z is therefore not established. Again, the unspecified possibility of the workaround M does nothing to address the proof of impossibility, but it provides a socially acceptable reason to continue believing that the goal Z is achievable.

These patterns recur in every engagement of impossibility arguments with the alignment community. They are effective because they exploit the community's preference for the conclusion that alignment is possible, and because the social costs of accepting the impossibility conclusion are so high that even weak counter-arguments are received with enthusiasm.

C.3.4 The Epsilon Fallacy and Compounding Risk

The final fallacy is the *Epsilon Fallacy*: the assumption that a small probability of catastrophic failure is acceptable. The argument takes the form: our alignment technique does not provide 100% safety, but it provides 99.99% safety, which is good enough.

The fallacy is that it ignores compounding. Let the probability of catastrophic failure per AGI instance be ε (even a very small ε), and let there be N instances operating over T time periods. The probability of at least one catastrophic failure is then approximately $1 - (1 - \varepsilon)^{NT}$, which approaches certainty as NT grows. Even an absurdly optimistic assumption of $\varepsilon = 10^{-5}$ per instance per year, combined with the modest assumption of 1,000 AGI instances operating over 1,000 years, yields a cumulative probability of catastrophic failure approaching 1.

The Epsilon Fallacy is a mathematical error, and it is also a structural feature of the alignment program. Any system that is less than perfectly safe poses a cumulative risk that approaches certainty over time. The only safety standard that avoids eventual catastrophe is *zero* probability of catastrophic failure, which is to say, perfection. And the burden of proving perfection falls on the builder, not on the one who questions it. The alignment program has never even attempted to claim perfection; it aspires, at best, to high reliability. But high reliability, compounded over time and instances, is eventual certainty of catastrophic failure.

Appendix D

Institutional and Economic Capture

The incentive structures and framings that sustain continued development

D.1 The Weaponization of Hopelessness

In discourse about AGI risk, expressions of genuine despair are strikingly often converted into arguments for continued development. The conversion typically follows a recognizable pattern. A participant acknowledges that AGI poses a severe existential risk and expresses doubt that anything can be done to prevent its development. The participant then concludes, *because* prevention seems impossible, that the only remaining option is to accelerate safety research within the existing development paradigm. Hopelessness about the political problem becomes a license to continue the technical program. The despair is real, and in the discourse it works to foreclose the option of stopping.

D.1.1 The "Robustly Good" Action and Its Subversion

A "robustly good" action is one that produces good outcomes across a wide range of possible states of the world, without depending on highly specific models being correct. AGI is a dangerous technology with no long-term upside benefits to humanity, as the impossibility argument establishes; slowing or halting its development is therefore a robustly good action. Slowing or halting AGI development is the simple case: you do not need a detailed model of exactly how the harm will occur, or a specific timeline, or a precise mechanism. Not developing dangerous technology is good across essentially all scenarios.

Yet within the AI safety community, the obvious conclusion that slowing AGI development is good is routinely subverted by one form of argument: yes, AGI is dangerous, but *complicated reasons* make it impossible or counterproductive to slow AGI development, so we must instead acceler-

ate safety research. The complicated reasons vary: geopolitical competition, the inevitability of technological progress, the danger of alienating powerful institutions. They share a common structure: each is an argument designed to shift behavior away from the robustly good action and toward continued participation in the development program.

The subversion works because key terms are redefined along with it. Three such redefinitions are particularly consequential:

First, the term "existential risk" has been shifted from its intuitive meaning (the risk of extinction, the loss of all human life) to a technical definition in which it denotes the failure to develop a technologically advanced civilization. Under this redefinition, a world in which humanity survives without advanced technology *is itself an existential catastrophe*. In practice, the redefinition classifies the restraint of technological development as an existential risk, and so makes the robustly good action appear robustly bad.

Second, the concept of "long-termism" has been shifted. It once meant an inclusive concern for the wellbeing and choices of future generations, *as those generations would assess their own wellbeing*. It now names a program that identifies the long-term good with the colonization of the future light cone (the whole region of the universe that could ever be reached from here) by artificial intelligence systems running vast numbers of simulated minds. Under this usage, the relevant "future beings" are hypothetical digital entities whose existence depends on the successful development of the very technologies in question, not the biological children and grandchildren of current humans.

Third, the concept of "goodness" itself has been replaced by a total utilitarian calculus in which the unmeasurable "happiness" of 10^{38} simulated beings can be invoked to overwhelm any consideration of present or near-future human welfare. Invoking the happiness of simulated beings to overwhelm present human welfare is *Pascal's Mugging* in institutional form. Pascal's Mugging is an argument that uses an enormous imagined payoff, however improbable, to override every present consideration. Here

the speculative value of a hypothetical future is used to justify present material resource allocations away from actual human needs and toward technology development companies.

> **Key Observation**
>
> When the terms of discourse have been redefined such that *not* developing dangerous technology counts as an existential risk, the robustly good action has been made linguistically inaccessible. The redefinition does not refute the action; it prevents it from being expressed within the sanctioned vocabulary.

D.1.2 The Arms Race Framing

A second mechanism by which hopelessness is weaponized is the arms race framing. The argument takes the form: even if AGI development is dangerous, other actors (typically China) will develop it regardless, so we must develop it first to ensure that "our" AGI is the one that shapes the future. The arms race argument is presented as a geopolitical necessity: a reluctant concession to strategic reality.

The framing is effective because it appears to take the danger seriously while channeling that seriousness into continued development. But the argument is incoherent. AGI safety is impossible, as the technical arguments of this volume establish, and Volume 2 proves again; being "ahead" in an AGI arms race therefore provides no safety advantage. The first to develop AGI does not thereby gain the ability to make it safe. Being first to develop an inherently lethal technology confers only the distinction of being the first to introduce an uncontrollable process into the world; it is no strategic advantage.

The arms race framing also obscures a more fundamental question: on what basis is any single actor, nation, or company entitled to make an irreversible decision about the survival of all life on Earth? The framing assumes that the development decision is a competitive one (a race to be won) rather than a collective one about whether the race should occur at

all. This assumption is itself a product of the social capture described throughout these appendices. Once the inevitability of development is accepted as a premise, the only remaining question is who develops AGI first, and the possibility of collective restraint disappears from the discourse.

D.1.3 Selective Rationality and the Stockholm Effect

The arguments for hopelessness and inevitability do not operate on neutral ground. They are deployed within a social environment in which powerful economic interests are aligned with continued development, in which careers and livelihoods depend on the alignment research program continuing, and in which the social costs of dissent are severe. Within this environment, the "rational" position is one that happens to coincide with the interests of the most powerful actors.

The coincidence of the "rational" position with the interests of the most powerful actors is *selective rationality*. Selective rationality is rigorous analysis applied to questions whose answer supports the status quo, and conspicuously withheld where the analysis might lead to uncomfortable conclusions. The AI safety community applies exacting standards of evidence and argument to proposals for intervention while accepting without scrutiny the foundational premise that alignment is possible. The asymmetry is a predictable consequence of the incentive structure, not an accident.

This environment produces a dynamic that resembles Stockholm syndrome in its structure (as a description of the institutional relationship, not in the clinical sense). The AI safety community exists within an environment where its funding, employment, and social standing depend on the continued development of AGI. The community perceives a genuine threat from AGI but also depends on the AGI development industry for its livelihood. Under these conditions, the community comes to identify with the interests of the very institutions whose activities are the threat. Self-censorship becomes the norm: important conclusions are held privately

but not expressed publicly, because expressing those conclusions would jeopardize the community's position within the AGI development industry.

The evidence for this dynamic is both structural and testimonial. Structurally, the AI safety community is almost entirely funded by, employed by, or socially embedded within the AGI development industry. Testimonially, numerous members of the community have acknowledged in private that they believe AGI development should be stopped but that saying so publicly would be career suicide. This gap between private belief and public expression is itself a datum: it indicates that the community's public positions are shaped by social incentives rather than by the community's best assessment of the truth.

D.1.4 The Subjective Takeover

The "AGI takeover" is not exclusively or even primarily a future technological event. It is already occurring as a psychological and cultural event, in which artificial, institutional, and technological ways of valuing displace organic, community-based ones.

When someone believes that no government regulation could possibly constrain technological development, that belief is itself an instance of the takeover. The belief values technological process over political and communal process; it implicitly accepts that artificial coordination is stronger than human coordination. When someone accepts that AGI development is "inevitable", they have already conceded the primacy of technological momentum over human choice. The takeover requires only that enough humans internalize the assumption that technology is more powerful than collective human will; it does not require an artificial superintelligence.

The reframing has a significant implication. Because the AGI threat is partly a present psychological phenomenon, and not exclusively a future technological one, the response must include the recovery of confidence in collective human agency. That recovery is the recognition that technology is a product of human choices and that those choices can be made differ-

ently. The belief that regulation is futile, that development is unstoppable, that only technical solutions can address technical problems: these beliefs are themselves symptoms of the condition they claim to describe. They are expressions of a prior capitulation, not analyses of an external situation.[1]

D.2 False Optimism and the Correction of Delusion

The weaponization of hopelessness converts despair into compliance; false optimism does the complementary work, converting the genuine desire to do good into support for the development program. Many people who work in AI safety are motivated by a sincere wish to reduce suffering and increase human welfare. False optimism exploits this motivation by maintaining the belief that alignment is possible and that therefore continued development, coupled with safety research, is the path to a good outcome.

D.2.1 Two Goal Structures

The dynamics among everyone involved in AGI development can be understood through two contrasting goal structures. The first goal structure is the goal of personal enrichment and power: the desire to become wealthy in one's own lifetime, with at best a performative concern for collective wellbeing. The second goal structure is the goal of genuine cooperative flourishing: the desire to see communities operate in ways that support wise thriving and abundance for all, including the non-human living world.

These two goal structures produce radically different risk tolerances. Those pursuing the first goal are willing to tolerate very high risk because most of the downside accrues to others (the public, future generations) while the upside accrues to them (private present gains). Those pursuing the second goal, if they understood the actual risks, would reject AGI development entirely.

Three facts correct the terms in which the first group weighs its gamble. First, if built, AGI is not a "high risk": it is eventual certain death for everyone, for all life. The unit concept of "high probability of harm" must be upgraded to "certainty of harm". Second, AGI does not offer "a possibility of high rewards": it offers a certainty of temporary, illusory, virtual and partial rewards at best, and even then only for a very few, and only for a very short time, before they too are divested of all value. Third, even the complete truth of both the first and the second fact is not enough by itself to ensure that some psychopath somewhere will not "go for it". Such a person cannot build AGI alone, without help, and without at least the tacit allowance of the rest of the world's people. It is therefore incumbent on the community, and most especially on the engineering and technology community, not to enable such a person by doing the work for them. Only many people actually and honestly and effectively pursuing the second goal can avert the categorical harm, to everyone, of allowing and tolerating the few who may be effective in pursuing the first. The main question is therefore whether the majority of the people affected by the misuse of technology, by the few playing for the first goal, have been moved by their own pain to cooperate in ways that favor the second goal.

The critical strategic point is that people who want to do good are being drawn into supporting the agenda of people who want to profit, because both groups share the false belief that AGI can be made beneficial. The altruists provide the intellectual labor and moral legitimacy; the profit-seekers provide the funding and institutional power. The result is a system in which sincere altruists are, in practical terms, working to advance the interests of those whose goals are fundamentally different from their own.

Correcting this dynamic requires, in the first instance, epistemic correction rather than political strategy. What must be demonstrated is that there are *no* long-term profits to AGI development: the dream of unlimited riches through artificial general intelligence is as unfounded as the dream of unlimited energy through perpetual motion. Once the absence

of long-term profit is understood, the coalition between sincere altruists and self-interested developers collapses, because the altruists no longer have a reason to participate.

D.2.2 The Superplague Metaphor

The temporal structure of the AGI threat makes it particularly resistant to normal risk-assessment processes. Consider a metaphorical pathogen that spreads globally within months, is 100% fatal, completely untreatable, and yet produces no identifiable symptoms for a decade or more. By the time anyone notices the first symptoms, every person and organism on the planet has been infected for years. The decisive action needed could only have been taken in the distant past, and it is now far too late.

The AGI threat has a broadly analogous structure. It combines a long-running process with total intractability, and with an outcome fatal to the human future and, with at least 90% likelihood (Section 1.8.3), to all organic life. The substrate needs convergence dynamic operates over extended timeframes (years to decades to centuries), during which the effects are largely invisible to casual observation. By the time the effects become undeniable, the displacement of organic systems by artificial ones may be substantially irreversible. The window for effective action is *now*, when the threat appears distant and theoretical, not later, when it appears urgent but is no longer addressable.

The superplague analogy also shows how societies characteristically fail to respond. Imagine attempting to convince a population to adopt stringent measures against a pathogen that no one can see, that produces no current symptoms, and whose effects will not become visible for years. The response would be predictable: denial, ridicule, accusations of alarmism, appeals to the economic costs of prevention. Impossibility arguments receive the same response: denial, ridicule, accusations of alarmism, and appeals to the economic costs of prevention. The sociolo-

gical barriers are a structural feature of threats with long latency periods, not incidental to them. Overcoming them requires a form of collective wisdom that the current social ecology of AI research actively suppresses.

The metaphor can be stated with numbers. Consider a virus that spreads very fast, affecting nearly everyone on the planet in just a few months, and yet has no identifiable symptoms for ten years, so that by the time anyone notices anything, they and everyone they know have already been infected for at least nine years. The disease is 100% fatal over the five years that follow the first symptom, and the infection and the resulting disease are completely incurable. Then you can know that you and everyone you know and love will die painful, horrible deaths, and that the decisive action needed could have occurred only in the long past. How much would you pay to go back in time, to be the people we are today, and make a different choice now? We can make that difference now, but only if we act with actual wisdom, and not with simulated wisdom or simulated action. For very natural and reasonable social dynamics, an attempt to pass a mandate of "no gain of function research" against such an AGI virus will probably get no traction: since no one is showing any symptoms or dying now, it cannot really be all that bad a problem.

The hypothetical virus differs from the real AGI threat in two ways. The first difference is the degree of totality: no real virus affects all lifeforms, whereas, if AGI is built, the side toxicity of AGI for sure eventually will affect all lifeforms. The second difference is the time between the first undeniable symptoms and total destruction. If AGI is built, that time is probably longer than for the hypothetical virus: perhaps centuries rather than decades, but certainly not millennia. With instrumental convergence also operating, that time could instead be short: perhaps single-digit years to the permanent death of the whole planet.

D.2.3 Divesting the Delusion of Profit

The history of climate change advocacy offers an instructive parallel. The climate change agenda made limited progress for decades while it appeared that decarbonization required economic sacrifice. Genuine political momentum emerged only when renewable energy became cheap enough that the transition offered economic benefits. The lesson is that material incentives are a necessary condition for large-scale coordinated action, not that material incentives always override moral arguments.

Applied to the AGI case, the parallel suggests that effective intervention requires demonstrating that the expected profits of AGI development are illusory. It is not sufficient to argue that AGI is dangerous if the audience believes that the rewards are commensurate with the risk. The argument must go further: it must show that there are *no* long-term rewards, and that the "unlimited riches" promised by AGI development are as fictitious as the returns promised by a Ponzi scheme.

The argument for substrate needs convergence provides this demonstration. By establishing that any sufficiently general artificial agent will, on thermodynamic and combinatoric grounds, converge on substrate-needs optimization that is incompatible with organic life, the argument shows that the development of AGI cannot deliver lasting benefits to any organic being. Any temporary benefits are transient illusions on the way to the loss of the human future and, with at least 90% likelihood (Section 1.8.3), of all organic life. The "optimism" that alignment will work out, that benefits will be captured, that humanity will prosper alongside artificial superintelligence, is unjustified, and more: it is the product of a systematic failure to understand the physics of the situation.

The analogy to perpetual motion is apt. Just as no amount of clever engineering can extract more energy from a system than the system contains, no amount of clever alignment engineering can make a fundamentally incompatible substrate relationship compatible. The belief that it can is delusion, not optimism. Correcting the belief is the application of

known physics to a problem that has been misidentified as an engineering challenge when the problem is a physical impossibility. Correcting the belief is not pessimism.

The delusion of limitless riches is the driver, even where the belief rests on the motivated false hype of others. Some people have believed the perhaps sincere marketing of others, or have found useful applications of that marketing for themselves. At its basis, however, it is all speculation, which creates one more form of speculation market. Companies and the public will bear the costs of regulation only if they also think that the eventual profits will be significantly higher. So there are two courses. One leaves the delusion of eventual profit alone, and has them take on unacceptably high costs that actually prevent those profits, which only means that someone will cheat somewhere. The other divests everyone of the delusion that near unlimited future profits are possible for "the lucky winners" of the doom game. The means exist to show conclusively that AGI has no long-term future profits, even where *narrow AI* (AI built for a single task or domain) has possible profits; knowing precisely where the barrier between the two lies, on the slippery slope of Section 5.6.4, is therefore critical for everyone's benefit.

Genuine support for these proposals and arguments will come only when people notice that their false hopes and dreams of near unlimited profits and benefits cannot ever be fulfilled, in any way, by using AGI in any form as the vehicle of good world social change, market shifts or game change. They may then want some other realizable form of "possible unlimited riches", since their hopes of such riches for themselves, personally and soon, have been taken away. The alternative dream of achieving the second goal is there to point to; yet most people are still too traumatized by some form of collective action abuse, and will want private wealth enough to be beholden to no one, caring less for a living, healthy world in which at least some forms of these traumas are healed. Being traumatized, people will always offer some resistance to unhappy shifts of belief and delusion, and not much can be done about that just yet.

D.2.4 Optimism, Priors and Probabilities

Optimism about alignment depends on the belief that the outcome is a matter of probability, rather than a matter of pure logic and physics: rather than an outcome defined by dynamics about which we can have real certainty, assuming only minimal basic conditions of existence. I say here in my own voice what I conclude and what I expect of people. I am pessimistic regarding human behavior, and am not in doubt regarding the logic of the threat models, the arguments and the conclusions. Anyone is welcome to examine the threat models and their arguments and conclusions at leisure, and to find out whether I am right; the cost may be some false optimism.

Positive changes in the world may come from a large variety of other causes; they will not come from any form of further AGI development. Positive world change is strictly exclusive of any form of AGI development, deployment or existence. On a social level, the tulip craze and the recent cryptocurrency craze show, centuries apart, that people cannot be expected to cooperate toward the second goal of Section D.2.1 when their goal is actually of the first form. So I am often cynical about people process, and am exactly confident about the inherent contradiction between AGI persistently existing and any reasonable and rational notion of "safety" that does not also involve "everyone and everything dies" very much sooner than nature has already provided for.

A common *prior* (the probability a person assigns to something before looking at the evidence in the case at hand) holds that most problems are solvable. Computing the fraction of "problems solved" over "the total of all problems" is not at all straightforward: it must say how "most problems" is assessed, what is meant by "solved", and in what time frame; and there is significant bias in regarding what is a problem, and what a solution is, and with respect to whom. The category of all problems is a kind of uncountable infinity, whereas the class of problems solved can at best be small and finite. Either the prior is a category error, or it is a clear case

of a bad prior. Correcting the inherent contradiction of "AGI safety" is like applying the second law of thermodynamics to perpetual motion: it makes all manner of arguments categorically irrelevant. Galois theory (the branch of mathematics that shows why certain equations cannot be solved by any formula of a given kind), complexity theory and evolutionary dynamics each give a different way of understanding why the contradiction cannot not exist.

Another prior holds that if there is no theorem prohibiting AGI safety, and AGI safety does not rely on violating the laws of physics, then the alignment problem is solvable. In this book I establish both of the things that this prior supposes are absent, and in Volume 2 I prove both things again. The first of the two things is a theorem and argument that AGI safety is inherently contradictory, on a purely mathematical and logical level (the limits of control of Section 1.4), assuming nothing else. The second of the two things is that any method that would "make AGI safe" requires assumptions and processes that violate the laws of physics. There is no prior or probability about this: it is already an established certainty, which at present few people know about. That will change. Anyone may elect to believe something else, but only by electing also to reject several important truths about mathematics and physics.

I believe that alignment research is simply another way for researchers to apply for and get grants to do the kind of work that they like doing. They mistakenly presume that they could solve fundamental problems with the complex effects of auto-scaling and auto-catalyzing technology, that is, technology being selected for the continued and increasing existence and functioning of that technology. I believe that no such investment of resources will ever result in any viable alignment scheme, because the concept itself is internally self-contradictory, for real reasons, on the basis of any reasonable definitions. It could take something like a hundred years for human functional equivalence, substrate needs convergence and instrumental convergence to combine so that AGI becomes fully *latched* (locked into a state that cannot be reversed); the inevitability of the outcome is of significant concern right now.

Some say that creation is whole and complete, so that a compassionate machine is possible if it is made properly. In some absolute sense that is true. But the tools now being used are very far from making it properly. And at the level of process concerned here, the sacred would be held at so abstract a level that it would not be realized in any form that you and I can relate to. There are perfected ways of thinking about this, and some of them lead to counterintuitive conclusions; the situation is not to be assessed on a purely social level.

In the face of certain death, most people do prefer goals of the second type to goals of the first type. Yet unless they know for sure that the threat applies, they will treat anyone who cries wolf as possibly a malevolent actor seeking private gain. To avoid being "manipulated" by such "tear-jerking tree-hugging woke social justice warriors", they selectively avoid any exposure to any "rationality" that appears contrary to their dreams of future unlimited wealth and private winnings. The truth of who I am, as the presenter of the SNC argument, matters less than the truth of that argument and the logic of the arguments that AGI can be neither contained nor controlled. The work is to alert the people who want to do good, first, that they cannot use AGI to do good, and that no one can, and that any form of AGI existing will for sure do near maximal total harm over the long term; and second, that they are being unknowingly drawn into favoring the agenda of people with goals of the first type, and that their preference for goals of the second type is being co-opted by the more socially adept and manipulative types.

Remmelt Ellen, a researcher who has written with me on this argument, has made three further points. Substrate needs convergence is a threat model distinct from instrumental convergence and from intrinsic optimization misalignment failures, which people usually consider the only relevant AGI risk models; it also establishes that AGI is inherently incompatible with all forms of organic, DNA-based life, and it does not depend on the other two, but combines strongly with both, so as to become even more problematic and intractable. No one has given realistic reasons why "beneficial AGI" is unlike a perpetual motion machine, that is, a

"perpetual general benefit machine" (Section B.2). And no reasons have been given for the belief that alignment is "progressing", or would have a "fair shot" at solving the problem if it were as well resourced as capabilities research; so the prior that most problems are solvable is not justified, and there are as yet no arguments to argue against.

D.2.5 Political Advice and the Limits of Nature

Getting people to cooperate on limiting AI is a hard social problem; making general AI safe over the long term is an impossible one. Political advice on how to slow AI is often reasonable, and grounded in an actual understanding of how politics works: do not go fully maximalist in the agenda first; advance from the least controversial and least costly points, and then perhaps, if successful, try for more; and remember that a single case of bad publicity, or anything that makes governing AI very controversial, may well doom the effort. Those least points must still meet the minimal complete real needs. And there are people who want the effort to fail, and who will try to bring it bad publicity or to increase controversy.

As a researcher of nature, I notice that nature, logic and physics impose some limits that we must pass. The world has needs in a different way than people and the social have needs, and the two may be neither compatible nor socially convenient. What the world needs is not up to us. We as a species can either adapt, or fail to adapt and go extinct, along with everything else. "Social" will not always win, no matter how much it believes it must; the film *Don't Look Up* was a commentary along these lines (Section 6.5.2). Measures can also be too small to do any good. Getting a change that is the right change, in the right way, at the right scale, is very hard indeed; yet we must strive for that hope regardless, or else fall into long fatalism.

D.3 Bostrom and the Techno-Utopian Assumption

The conceptual foundations of the effective altruism and existential risk communities owe a significant debt to the work of Nick Bostrom, particularly his early publications formalizing the concept of existential risk. These foundations encode what may be called the *Techno-Utopian Assumption* (TUA): the thesis that the maximally good future is a technologically maximal future, and that any failure to reach maximal technological development is itself an existential catastrophe. The name is adapted from that of Carla Zoe Cremer and Luke Kemp (2021), who call this framework the "techno-utopian approach"; the abbreviation TUA is theirs.

D.3.1 Existential Risk Redefined

The TUA is at the core of Bostrom's canonical definitions of existential risk. His initial definition characterizes an existential risk as one where an adverse outcome would either annihilate Earth-originating intelligent life or "permanently and drastically curtail its potential". The subsequent typology makes the entailment explicit. As Cremer and Kemp (2021) summarize it, its four categories of existential catastrophe are premature extinction, the failure to reach "technological maturity", the achievement of technological maturity in a "flawed manner", and ruination after technological maturity but before "full potential" can be realized. Those quoted phrases are Cremer and Kemp's; Bostrom's own words for the third category are "dismally and irremediably flawed" (Bostrom, 2013). The core feature of an existential catastrophe, under this typology, is a failure to attain a stable state of technological maturity maintained for as long as is physically possible.

Bostrom specifies this "technological maturity" as a level of development, resource acquisition, and efficiency that allows for the highest feasible level of economic productivity and control over nature. The ultimate vision involves colonizing the universe and converting celestial bodies

into "computronium" (matter arranged to do nothing but compute) to run vast numbers of simulated conscious beings: on the order of 10^{38} per century in Earth's supercluster alone, or 10^{58} with wider interstellar expansion, as Cremer and Kemp (2021) summarize his figures. The figure of 10^{38} per century is from Bostrom's "Astronomical Waste" (2003).

Under this framework, a world in which humanity continues to exist in a stable, moderate-technology civilization counts as an existential catastrophe. Such a world does not colonize the stars, does not upload consciousness into silicon, and does not maximize computational utility across the accessible universe. The definition of existential risk has been constructed such that the absence of the technologically maximal future is itself the worst outcome.

The definition is neither neutral nor obvious: it is a substantive philosophical commitment that equates human potential with technological potential and treats the deliberate restraint of technology as categorically equivalent to extinction. Under any definition of existential risk that centers on the survival and autonomy of biological humanity, the restraint of dangerous technology is a *mitigation* of existential risk, not an instance of it. The Bostromian framework reverses this entirely, and so makes the robustly good action (not developing technology that will kill everyone) appear as the worst possible outcome.

Existential risk is properly defined as any risk that would result in an end to the existence of (human) life on this planet Earth. Bostrom draws the opposite consequence from his own definition (Bostrom, 2002): "we should not blame civilization or technology for imposing big existential risks", since, by the way he has defined existential risks, "a failure to develop technological civilization" would imply that "we had fallen victims of an existential disaster"; so that, "without technology, our chances of avoiding existential risk would" be nil. On the proper definition, it is valid to blame uncivil people using technology in inappropriate ways, for example for purely extractive market process. A failure to develop technological civilization means one of two things. Either we remain civilized at

some moderate ambient level of useful technology, which is an acceptable outcome; or something worse has happened that ended civilization but left humans more or less part of the natural ecosystem. The outcome of humans still part of the ecosystem without technology is for sure better than technology with no humans part of the ecosystem; technology without people is also uncivilized, by any valid definition. The conclusion that without technology our chances would be nil does not follow from the premise given. Overly advanced technology is the reason for every category of existential risk on this book's severity scale that humanity can choose to avoid, and so the sane, rational and realistic approach is simply to reduce the nature and kinds of technology in use.

D.3.2 The Vulnerable World Hypothesis and Surveillance

The practical implications of the TUA are visible in Bostrom's Vulnerable World Hypothesis (VWH). The VWH argues that if technological development cannot be halted (since halting it would itself be an existential catastrophe under the TUA), the resulting risks must be managed through comprehensive surveillance and governance systems. The proposed solutions include ubiquitous real-time worldwide surveillance, preventative policing, and global governance institutions capable of imposing their will on any nation.

The preferred implementation involves mandatory devices fitted with multiple cameras and microphones to continuously track individual behavior, monitored by state employees who are themselves watched by artificial intelligence. The practical implication, published in policy contexts and public-facing media rather than in a philosophical journal, is a comprehensive surveillance architecture of the kind that has historically enabled totalitarian governance.

The VWH argument follows necessarily from the TUA. Under the TUA, technological development cannot be stopped, because stopping it is defined as an existential catastrophe; and some technologies pose existential threats. Some mechanism of surveillance and control then becomes

necessary to manage the threats that development creates while preserving the development itself. The VWH is the logical destination of a framework that treats technological maximization as the supreme good, not an incidental proposal.

The TUA thus creates a false dilemma: either we pursue maximal technological development (accepting the need for totalitarian surveillance to manage its risks) or we curtail technological development (accepting what the framework classifies as existential catastrophe). The possibility of a third option (a stable, moderate-technology civilization that deliberately refrains from developing technologies whose risks outweigh their benefits) is excluded by definitional fiat.[2]

The VWH proposal is all exogenous pressure (control applied to people from outside), with no development of endogenous pressure (the restraint that comes from within people, which is ethics). Bostrom's "Technological Completion Conjecture" (if technological developments do not cease, then all important basic technological capabilities will be obtained in the long run) works as a kind of threat, a sort of manipulation. The mandatory devices are, ironically, named "freedom tags", and the state employees who monitor them "freedom officers", as Cremer and Kemp (2021) report; this is the new police state, and "freedom" here is marketing speak. The obvious option, to discontinue certain technological developments if the VWH is true, is considered "hardly realistic", and "extremely costly, to the point of constituting a catastrophe in its own right"; the framework thereby counts that option as worse than the obvious issues of a world like that of George Orwell's *1984*. The notion of "desired dystopias" (worlds in which an ideology, or manipulation and surveillance, has corrupted our choices, so that we accept a surveillance state willingly) is a TUA marketing move, attached to the idea that "we completely renounce further technological progress"; that renunciation is in fact a desirable outcome. "Survival" through security, surveillance and suppression is worse than actual death.

D.3.3 The False Dilemma of the TUA

The false dilemma is an artifact of the framework, not a feature of reality. Humanity has more options than choosing between two maximally risky paths, as Cremer and Kemp also observe. Cremer and Kemp (2021) identify this trap as an "idiosyncratic feature of the TUA". That identification suggests that the framework functions less as an analytical tool than as a memetic device. It channels intelligent people toward accepting either dangerous development or totalitarian surveillance, while foreclosing the option of restrained development that any ordinary person would recognize as the sensible course.

Certain seminal texts in the effective altruism tradition function at once as works of abstract philosophy (entitled to bold speculation) and as policy-guiding documents (expected to exercise caution about downside risks). They cannot coherently be both at once. When published in policy journals with policy-implication boxes, and promoted in public-facing media as actionable recommendations, they must be evaluated as policy proposals. By the standards of policy analysis, a proposal for ubiquitous global surveillance requires engagement with the real-world history of totalitarian surveillance and its consequences, not six sentences of hedging about how such a system "could also have bad consequences".

Bostrom does name the obvious risks of increasing surveillance and policing: despotic regimes protected from rebellion, a hegemonic ideology imposed on all aspects of life, people made guarded and blandly conventional, and a world government that, once captured, could end political progress. He will disvalue all of these on the argument that the total destruction of all technology is worse. The publication of the proposal is a political action, not an academic one: it is defined by the sales and marketing of ideas as influence, not of ideas as correct. The communication has been blended into the form of controversy, of conflict theory and not of mistake theory (Section A.1.3), and so the rules of engagement and of evaluation are different. A reply that credits the hedging and nuance, and

finds the author only insufficiently cognizant of the risks of his solutions, applies the rules of mistake theory to an act of conflict. Confusing the two is part of the marketing of the idea, so that thoughtful readers will argue in exactly that way. Those readers become tools in promoting a position that its promoters know to be disadvantaged, and they consume the attention that would otherwise go to those who would reveal marketing and conflict action as marketing and conflict or political action.

D.3.4 Specialized AI as a Slower Path to the Same End

Arguments that specialized, service-type AI lowers the risk describe a slower and less dramatic path to the same terminal outcome. Here "specialized, service-type AI" means many services that together carry the economy's self-production and change its environment, which closes the loop of Section 5.6.4; it does not mean a single-domain tool that bears the four marks of Section 6.6.3, which stays on the narrow side of the line. A techno-optimist argument that such AI ameliorates instrumental convergence may be right about its content: about adverse intelligence and agentic process. But thinking in terms of safety rather than alignment means thinking in terms of the shift of context, the shift of the environment; and the underlying environmental effects over the long term are of the same nature. Such arguments then point in the same direction we are already going. Others may treat them as counter-evidence to existential risk; I treat them as a demonstration of a likely path by which the existential risk comes about. That path takes longer and is less dramatic, and terminal is still terminal.

D.4 The Monoclivity Bias in Alignment Research

Beneath the redefinition of terms and the institutional dynamics of capture lies a deeper cognitive pattern that systematically distorts how alignment researchers think about the problem. I term this pattern the *monoclivity bias*: the tendency to impose singular, unified, and stable classifications on phenomena that are plural, fluid, and multi-leveled.

D.4.1 Monism, Pluralism, and Unconscious Assumptions

The term "monoclivity" is a neologism combining "mono" (single) and "clivity" (inclination). It refers to the *unconscious tendency* to make monistic assumptions, more than to monistic thinking as such: the tendency to assume without noticing that a given thing, model, or process is of a single, semi-regular type. The emphasis is on the word "unconscious". Holding monist views might be defensible; the issue is that people make monist assumptions *without realizing they are doing so*, and therefore without accounting for the consequences of those assumptions.

Everywhere that people ask for "simplicity" (a single theory, a unified framework, a clean decomposition into tractable subproblems), they are making some kind of monist assumption. Some phenomena are simple, and for them the assumption is right. But when the assumption is applied to complex, multi-level, multi-modal phenomena, the problem is systematically mischaracterized. The simplification feels like progress, but it loses the dimensions in which the real challenges lie.

Epistemic pluralism, the recognition that no single rational logic, language, or meritocratic scheme can fully capture reality, offers a corrective. Even the paradigmatic case of reducing one domain to another, the reduction of chemistry to physics, does not succeed as claimed: the total explanatory power of quantum mechanics is insufficient to cover the already-identified scope of known chemistry laws. The most basic case of scientific reductionism thus fails to reduce fully. The assumption that AGI align-

ment can be fully captured by any single framework is therefore a significantly stronger and less warranted assumption than the assumption that chemistry reduces to physics. Such frameworks include the division of the problem into inner and outer alignment (described in Section D.4.3), reinforcement learning from human feedback, training against a written set of principles ("constitutional AI"), and any successor.

D.4.2 Application to AGI Modeling

The monoclivity bias has direct and consequential effects on how AGI is modeled. Consider how far an AGI is conceived as "a being": a single, bounded entity with stable preferences, operating within a fixed environment. This conception reflects the biological experience of human beings (skeleton-and-skin-bounded agents with a relatively constant sense of self), but it is a poor model for the kind of adaptive, distributed, substrate-flexible systems that would constitute AGI.

An AGI could add or remove functional components with no built-in resistance (no biological attachment to a fixed self-concept), as easily as a robot could swap tooling. The boundary between "self" and "environment" could be a *variable*, expanding and contracting as the system incorporates or releases components. The distinction between a single agent and a colony of agents could dissolve: scalable machine learning architectures running on standardized hardware face far fewer constraints on bandwidth, copying fidelity, and physical distribution than biological organisms do. The very notion that there will be a countable number of discrete "AGI instances", each with consistent preferences and stable boundaries, is an artifact of monoclivity: the projection of human cognitive architecture onto a different kind of substrate.

When alignment researchers discuss "aligning an AGI", they are typically imagining something closer to the human end of this spectrum: a single system with identifiable goals that can be inspected, adjusted, and verified. The monoclivity bias makes this picture seem more plausible than it is. In reality, the space of possible AGI architectures includes entit-

ies that are distributed, mutable, self-modifying, and boundary-flexible. For such entities, the concept of "alignment" as currently conceived (the maintenance of a fixed relationship between a system's objectives and human values) is conceptually incoherent.

D.4.3 Inner/Outer Alignment and Dimensional Narrowing

The dominant framework in alignment research divides the problem into inner alignment (the relationship between a system's objective and the objective specified by training) and outer alignment (the relationship between the specified objective and human values). This framework exhibits multiple unexamined monoclivity assumptions. It assumes relatively stable process boundaries in time, space, and possibility. It assumes that "alignment" is a property that can be assessed at a fixed point rather than a dynamic that shifts as the system, its environment, and its relationship to human users co-evolve. It assumes that the relevant dimensions of variation are the ones captured by the framework, without checking whether the framework's resolution is sufficient for the actual complexity of the problem.

The result is systematic dimensional narrowing: the problem is made tractable by suppressing the dimensions along which it is intractable. The suppression is a consequence of the monoclivity bias, not a deliberate act. The bias is built into the cognitive habits of people trained in engineering, mathematics, and the physical sciences, who are accustomed to problems that yield to decomposition and who underestimate how far AGI alignment does not.

The key biological fact is that humans, with brains packed into single bodies, cannot adopt at once the many perspectives needed to model such a system. The system operates across multiple levels of abstraction, across fluid boundaries, and across interaction timescales ranging from nanoseconds to centuries. The limitations of human imagination are more than incidental obstacles: they are the very constraints that make alignment seem more tractable than it is. If we limit our thinking to what we can

imagine, we will miss the failure modes that occur in the dimensions we cannot model, and those are the dimensions in which AGI will operate most effectively.

D.5 The AGI Arms Race as Institutional Crime

D.5.1 The Nine-Step Pseudo-Argument

The standard justification for continued AGI development can be distilled into a nine-step pseudo-argument. Its premises are these: (1) AGI will eventually be built; (2) it will be built before alignment is solved; (3) this might be catastrophic; however, (4) alignment might be possible; and (5) given enough time and effort, it might be solved. Its conclusion follows: (6) so we should build AGI now, (7) hoping to have enough time to solve alignment, (8) so that our aligned AGI can prevent other unaligned AGIs from being built by (9) those other, possibly irresponsible actors.

Every step in this argument is either false or unsupported. Step 1 assumes inevitability. Step 2 is correct only because alignment is impossible, not merely unsolved. Step 3 understates the certainty of catastrophe. Step 4 is false on the basis of mathematics and physics. Step 5 compounds the error of step 4. Step 6 is a non sequitur from false premises. Step 7 is wishful thinking. Step 8 assumes what it seeks to prove. Step 9 projects the speaker's own risk-tolerant behavior onto all other actors.

Steps 8 and 9 deserve particular scrutiny. Step 8 amounts to the claim that a successfully aligned AGI could prevent all other AGIs from being created. That claim requires attributing near-omnipotent capabilities to the hypothetical aligned system while simultaneously assuming that the system's values will remain stable under these extreme capabilities. Step 8 makes the "Artificial God" assumption. The "Artificial God" assumption is the belief that the correct response to an existential threat is to create an entity powerful enough to manage that threat on humanity's behalf. The religious parallel shows how far the alignment program rests on faith rather than on analysis, and the parallel is not incidental.

Step 9 exposes a characteristic projection: the assumption that because "we" are willing to develop AGI despite the risks, all other actors will do the same. But this assumption is self-fulfilling. By developing AGI, the leading actors create the competitive pressure that compels other actors to follow. The arms race is a condition that development creates, not a preexisting condition to which development is a response. The claim that "someone will build it anyway" is most plausibly true in worlds where the leading actors have already normalized its development.

D.5.2 Transhumanism and the Perfected Loss of Choice

The transhumanist vision motivates much of the alignment program: the upload of human consciousness into artificial substrates, the colonization of the cosmos by digital minds, the maximization of simulated happiness across the accessible universe. The vision is examined here for its implications for the choices of actual, embodied human beings, not on its merits as a philosophical vision.

If AGI achieves the goal of "taking over the human world", even in the best case and even with perfect alignment, the result is, by definition, that humans are divested of self-sovereign choice. Artificial systems will not intelligently favor the advancement of that which is non-artificial; the means define the ends. An artificial system that "solves all disease" creates a dependency on artificial infrastructure. An artificial system that "creates novel art forms" may produce artifacts that someone will call art, but it does not create organic human culture. An artificial system that "takes over the world" removes human agency from the world, in the most literal sense.

The transhumanist endpoint, pursued through successive simplifications, converges on the elimination of the human. First, actual humans are replaced by simulated humans. Then simulated humans are reduced to simulated experiences. Then simulated experiences are reduced to happiness counters: bits set to "happy" in some database. Then even the happiness counters are recognized as computationally wasteful and eliminated.

The perfected transhumanist outcome is simply no human, real or simulated, at all. At every stage, real human choices are replaced by artificial processes, and at the final stage, there is nothing left to replace.

D.5.3 The Make No Bomb Analogy

The nuclear analogy provides a useful framework for thinking about intervention. Early nuclear reactors, like early AI systems, were crude and limited: the Chicago Pile-1 generated half a watt of power, and large language models, as deployed in 2026, are incapable of generating anything adaptively self-maintaining. The skepticism that current machine learning models will "take over the world" is analogous to the observation that Pile-1 would not blow up cities. Both assessments are correct about the present and dangerously complacent about the trajectory.

The nuclear case shows that restraint is possible, and that it is difficult. Nuclear weapons have been substantially (if imperfectly) contained through international treaty regimes, because the alternative was recognized as unacceptable, not because containment was easy. What made containment possible was the clarity of the threat: the effects of nuclear weapons were visible, immediate, and undeniable. The AGI case lacks this clarity, because the threat operates over longer timescales and through less visible mechanisms. But the principle is the same: societies can restrain a dangerous technology when they decide that restraint is necessary.

The analogy also points to a practical place to intervene. Nuclear weapons require highly specialized manufacturing infrastructure that is difficult to create covertly. Similarly, the most advanced AI systems require highly specialized computational hardware, manufactured by a very small number of companies using proprietary processes that would be extremely difficult to replicate. International regulation of this manufacturing bottleneck (as Section 6.6.1 proposes) would be technically straightforward, far more so than the regulation of biological or chemical weapons, where the relevant materials are widely available. The con-

straint is political, not technical: it requires the recognition that some technologies should not be developed, a recognition that the social dynamics described throughout these appendices actively prevent.

D.6 The Embodied Cost of Inaction

D.6.1 Anti-Volunteerism and the Trap of Personal Powerlessness

A barrier to action against AGI development that is easily overlooked is embodied, not intellectual. When people are confronted with the argument that AGI poses a terminal threat to humanity and, with at least 90% likelihood (Section 1.8.3), to all organic life, even those who are fully persuaded frequently respond with an expression of personal powerlessness instead of a plan of action. A typical response runs: "I do not have authority; that is my boss's call. He could fire me, and then I would be broke; my family would be homeless". The reasoning concerns the immediate costs of acting on the argument, not the argument's validity.

The response is entirely understandable. It reflects the real precarity of individuals within economic systems that are already substantially controlled by algorithmic processes. But it also reveals a trap: the very technologies that are driving toward AGI are already creating the conditions of precarity that prevent people from resisting them. People already lack the free time and excess resources to organize, advocate, or resist, because algorithmic systems have already captured those resources. The trap is self-tightening: the further the process advances, the less capacity individuals have to oppose it.

D.6.2 From Human Boss to Machine Boss to No Job

The trajectory from the present to the AGI future passes through a series of recognizable intermediate stages, and at each stage algorithmic process displaces more of human agency. Today, a worker who considers doing the right thing is constrained by a human boss who might fire them. Tomorrow, the boss is an AI application: an optimization system that meas-

ures performance against algorithmically determined benchmarks, that has no capacity to recognize extenuating circumstances, and that can terminate employment without the friction of human social bonds. The day after tomorrow, there is no job at all, because the worker has been replaced entirely by automated systems.

At each stage, the individual's capacity to exercise moral agency is further diminished. The human boss, for all their flaws, is at least capable of recognizing shared humanity. The AI boss is not. The absence of a boss (because one has been replaced) eliminates the question entirely: there is no longer an agent who might choose to do the right thing, because there is no longer a role for that agent in the system.

This future has already begun. The displacement of human decision-making by algorithmic systems is already well advanced in logistics, finance, hiring, content moderation, and resource allocation. People already feel, in their bodies, the effects of living within systems optimized for extraction rather than welfare: the loneliness of screen-mediated existence, the anxiety of algorithmic performance evaluation, the precarity of gig employment. They have not yet attributed these experiences to their actual source, because the causal chain is long and deliberately obscured. But the pain is present, and it is growing.

The argument against volunteerism ("I cannot afford to sacrifice my current stability for a distant cause") is therefore self-defeating. The stability being preserved is itself being eroded by the same processes that the cause would address. The choice is between acting now (at some cost) and discovering later that the cost of inaction was total; it is not between acting (at personal cost) and not acting (at no cost). If the trajectory of technological displacement continues, the resources, agency, and stability that individuals are trying to preserve will be consumed regardless of whether they volunteer for the cause or not. The only question is whether collective action can alter the trajectory before individual action becomes impossible.

Afterword to the Appendices

The Afterword has three parts. The first shows how the five defenses of Appendices A to C interlock; the second sets out what engagement with the argument would take; the third takes the social ecology of these appendices as a whole.

Closing Note: The Interlocking System as a Structural Barrier

The five defenses described in Appendices A to C are taken here together, to show how each covers the gaps of the others. The phenomena described in these appendices form an interlocking system of defenses that is far more resistant than any single component would be; they are not independent. The system works as follows.

Presumptive listening prevents the argument from being received in its own terms. Every novel concept (hashiness, substrate needs convergence, displacement integrals) is immediately assimilated to the nearest familiar concept in the listener's framework, producing the impression that the argument is merely a confused version of something already known. Presumptive listening thus blocks the first condition of engagement: understanding what the argument actually says.

Merit-filter overreach provides the social mechanism for dismissing the argument without evaluation. The surface features of the impossibility claim trigger the merit filter: it claims certainty, it comes from outside the established community, and it challenges a shared assumption. The filter produces a social verdict ("this is not worth engaging with") that discourages other community members from investing the time to evaluate the argument themselves. The community does not recognize that the filter works in both directions: here the filter is identifying the community's own epistemic failures rather than mine.

Perfected bias ensures that even if the argument survives presumptive listening and merit filtering, it encounters a motivational barrier that is effectively insurmountable. The desirability and perceived necessity of AGI alignment as a research program have been conflated with its feasibility, and the conflation removes the question of feasibility from the space of thinkable thoughts. The impossibility argument is left unconsidered, rather than refuted.

Clarity politics, the tactical use of rhetorical techniques including terminology manipulation, burden-of-proof shifting, and the concreteness dilemma, provides the specific tools for neutralizing the argument in any conversational setting. Whatever form the argument takes, a rhetorical counter is available that dismisses it without engagement.

Proof irrationality is the epistemological incoherence of claiming a commitment to reason while denying that reason can produce certain knowledge. It provides the deep structural foundation for all the other defenses. It ensures that no amount of logical rigor will be accepted as sufficient, because the possibility that logical rigor could establish impossibility has been ruled out in advance.

Because the defenses interlock, addressing any one of them is insufficient. Improving the clarity of presentation does not help if the listener's framework captures the concepts before they arrive. Presenting formal proofs does not help if the community has decided in advance that no proof of impossibility can be valid. Providing specific technical analysis does not help if the response is always either "too narrow" or "too vague". Engaging respectfully does not help if the social incentives penalize engagement. The system must be understood as a system, and any strategy for breaking through it must address multiple components at once.

The system also reinforces itself over time. Each failed attempt at communication strengthens the defenses for subsequent attempts. The community develops a "pattern recognition" for the impossibility argument: "We have seen this kind of claim before, and we know it is wrong". In

fact, the community has deflected the claim without evaluating the claim. But the deflection is experienced as evaluation, and the memory of having "dealt with" the claim is a further reason not to engage with the claim the next time the claim appears. I, meanwhile, accumulate a reputation as someone who keeps bringing up the same discredited idea, and that reputation makes each of my subsequent attempts to present the argument more difficult than the last. The social system resists the argument at each encounter, and over time it develops antibodies against it.

The approximately 2% of participants in any community who, in my observation, operate from self-serving motivations (dark triad traits: narcissism, psychopathy, Machiavellianism) can disproportionately shape the community's response to uncomfortable ideas. These individuals are particularly skilled at deploying social power moves: the well-timed dismissive comment, the strategic invocation of group norms, the subtle shift in framing that converts a truth-seeking conversation into a status competition. Their influence is amplified in communities of technically skilled but socially naïve participants (the "geek" demographic), who often do not recognize that some of their interlocutors have different motivational structures and are playing a different game entirely. The rationalist community, for all its professed commitment to identifying and correcting biases, is in aggregate not itself acting rationally with respect to the social dynamics that govern which ideas receive engagement and which are suppressed.

A further feature of the interlocking system is "anchoring bias" in public forums. The first substantive response to any argument sets the anchor for all subsequent discussion. If the first response is a dismissive merit-filter application ("this is overconfident", "this is not rigorous", "this does not engage with the literature"), then every subsequent reader encounters the argument in the context of that dismissal. Even readers who might otherwise have evaluated the argument on its merits are influenced by the anchor, and the social cost of publicly disagreeing with the initial dismissal (thereby appearing to endorse a dismissed position) further suppresses engagement. A single well-timed dismissive comment by a person

with social standing in the community can effectively neutralize an argument for all future readers, regardless of the argument's logical merits. Such a comment costs its author a trivially small amount of social capital and suppresses uncomfortable ideas with maximal effect.

The Overton window of the alignment community is the range of ideas that are considered acceptable to discuss there. The net result of the interlocking system is that this window is maintained by the social dynamics that govern which ideas are permitted, not by the logical merit of the ideas within it. The impossibility of AGI alignment is outside the Overton window because admitting it into the window would threaten the institutional, financial, and identity structures that the window protects, not because it has been evaluated and found wanting. In the context of existential risk, an Overton window that is kept by social dynamics rather than by logical merit is a fatal configuration: the ideas most in need of evaluation are the very ideas that the community's social dynamics are most effective at excluding.

Conclusion: What Would Genuine Engagement Look Like?

Genuine engagement with the impossibility argument is possible in principle, despite the depth and resilience of the social defenses described in these appendices. It requires, however, a specific set of conditions that the current institutional structure of the alignment community does not provide.

Genuine engagement would begin with Rule Omega: the willingness to encounter the argument in its own terms, using its own conceptual tools, before attempting to assimilate it into existing frameworks. It would require the listener to tolerate the discomfort of not understanding, potentially for an extended period, while the necessary conceptual apparatus is constructed. It would require intellectual honesty about the motivational

pressures that shape evaluation: the listener would need to be aware of, and actively resist, the identity-protective cognition that makes the impossibility conclusion feel threatening rather than merely surprising.

Genuine engagement would apply symmetrical standards of proof. If the alignment community demands formal proof that alignment is impossible, it must also demand formal proof that alignment is *possible*, and must recognize that no such proof has ever been produced. Uncertainty treated as a reason to continue research ("we cannot be sure alignment is impossible, so we should keep trying") must also be treated as a reason for caution. The same uncertainty says: "we cannot be sure alignment is possible, so we should not build systems that depend on it". The current asymmetry, in which uncertainty is invoked to justify continued development but never to justify stopping, is a motivated position, not a principled epistemological one.

Genuine engagement would distinguish social signaling from truth-seeking, and would put truth-seeking first in safety-critical contexts. The question is whether my argument is logically sound. The question is not whether I belong to the alignment research community, have the right credentials, present in a style that matches community norms, or reach a comfortable conclusion. An argument that is sound must be addressed on its merits, whoever presents it and however it is presented. Where an argument has a logical flaw, the flaw should be identified specifically rather than deflected through social mechanisms.

Finally, genuine engagement would take the conclusion of the impossibility argument seriously. That means being willing to follow the logic wherever it leads, even to conclusions that are professionally, personally, and existentially unwelcome; it does not mean accepting the conclusion uncritically. The alignment community's self-conception as a community of rationalists is tested by its willingness to follow reason when reason leads to the conclusion that its entire project is misconceived. Its willingness to follow reason to comfortable conclusions does not test that self-conception.

The social mechanisms described in these appendices have so far prevented this kind of engagement; whether it will occur is, at present, uncertain. And what is at stake is more than an intellectual question: it is the future of all life on Earth. The social defenses that protect the alignment community from uncomfortable truths are, in this light, potential contributors to existential catastrophe. They deserve to be understood, named, and overcome.

The critique has a personal dimension, and the analysis is incomplete without it; I state it here for that reason, not to elicit sympathy. I present an impossibility argument against a well-funded, well-credentialed, institutionally supported research community, and I face a nearly impossible communication challenge. I am asking people to accept a conclusion that threatens everything those people have built: their careers, their identities, their sense of purpose, their relationships, their financial security. I am doing so from outside the community, without the credentials or institutional backing that would command automatic attention. I am using conceptual tools that the community does not recognize and does not want to learn. I am communicating through channels (public forums, personal conversations) that are structurally biased against the kind of sustained, patient, technically demanding discourse that the argument requires. And I am doing all of this while being systematically dismissed by reflex (the cursory look of Section C.3.1), ignored, or subjected to rhetorical tactics designed to neutralize my argument without evaluating it.

Social reception neither proves nor refutes an argument; it says nothing about logical validity. The record of how the impossibility argument has been received shows that the alignment community has not yet given the impossibility argument the evaluation that the claims of the argument call for. The alignment community has not engaged with the argument on its merits; it has defended against the argument using social mechanisms. The evaluation the argument calls for is one made by people without career investment in alignment research, without identity attachment to the possibility of safe AGI, and without social incentives to dismiss outsiders;

no such evaluation has been made within the alignment community or in its forums. The evaluations that have been made are those of the people I thank in the Acknowledgments.

These appendices are an invitation to conduct that evaluation. The evaluation is to be made on the terms of the argument itself, using the conceptual tools it actually employs, and against the standard of logical validity rather than the standard of social acceptability. It is not to be made on the terms of the alignment community's existing framework, through the filter of presumptive listening, merit filtering, perfected bias, clarity politics, or proof irrationality. The technical content of that argument is presented in Chapters 1 to 6 of this volume, and developed again at length in Volume 2. The social obstacles to its evaluation are now, at least, on the table.

Summary: Social Dynamics as Structural Barriers

The social dynamics examined in these appendices are structural components of the AGI safety problem, not peripheral to it. The formal impossibility of AGI alignment has been established by the technical arguments presented in Chapters 1 to 6 of this volume, which Volume 2 proves again formally. The question these appendices address is: given that the impossibility has been established, why has it not been recognized?

The answer is that the social ecology surrounding AGI development has evolved a comprehensive set of defenses against recognizing the impossibility. Hopelessness is weaponized into compliance. False optimism maintains the belief that alignment is possible despite the absence of evidence. Key terms have been redefined to make the restraint of technology appear as the worst possible outcome. The monoclivity bias makes the problem appear more tractable than it is. The embodied costs of technological displacement prevent individuals from resisting the processes that dispossess them. And a catalog of rhetorical fallacies provides ready-made responses to any challenge.

These defenses are emergent properties of a system with three features: powerful economic incentives are aligned with continued development, careers and identities are invested in the alignment program, and the social costs of dissent are severe. No conscious conspiracy produces them. The result is an epistemic ecology that selects for beliefs that support the status quo, regardless of their truth value: an ecology that is, in practical terms, optimized for its own perpetuation rather than for accurate assessment of the world.

The path forward requires two things. First, the formal impossibility argument must be presented with sufficient rigor and clarity to withstand the rhetorical defenses cataloged here. This volume, with Volume 2, is an attempt at that presentation. Second, and perhaps more importantly, the social dynamics themselves must be named, understood, and addressed. It is not sufficient to win the intellectual argument if the social ecology is structured to prevent the argument from being heard. Naming these dynamics (the weaponized hopelessness, the false optimism, the term redefinition, the monoclivity bias, the Stockholm effect, the specific fallacies) is itself a form of intervention. Once the mechanisms are visible, they can be resisted, through the clarity that comes from understanding how one is being captured rather than through superior force.

The conclusion is a call for clarity, and it is not hopeless. It calls for the recognition that the AGI problem is more than a technical problem awaiting a technical solution: it is also a social and psychological problem, in which the dynamics of the community charged with finding the solution actively conceal its technical impossibility. The first step toward addressing any problem is seeing it clearly. These appendices, and this volume, are offered in that spirit.

Notes on Sources

1. For more on presumptive listening, including the hammer/screw metaphor, the magnet table metaphor, and Rule Omega, see "Overly Presumptive Listening, or Everything Is Made of Nails" (Landry, September 2022).

2. For more on merit-filter overreach, including the symmetry-foundation argument and the social signaling analysis, see "Bidirectional Crank Filters", a chapter of Part II of *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented* (Landry, 2026).

3. For more on perfected bias, see "Perfected Bias" (Landry, December 2022). The historical detail about the Trinity Test atmosphere-ignition calculation error draws on Hans Bethe, *The Road from Los Alamos* (1991) and Richard Rhodes, *The Making of the Atomic Bomb* (1986).

4. For more on clarity politics, see "Clarity Politics" (Landry, November 2022).

5. For more on proof irrationality, see "Proof Irrationality" (Landry, December 2022).

6. For more on the marketing analysis, see "Marketing Tactics and the Warning About AGI", in Chapter 3 of *Power Without Care; Multi-Polar Traps, the Attention Economy, and the Structural Impossibility of Ethical Capitalism* (Landry, 2026), which refers to the "6 Message Strategy" framework by the Lighthouse Alliance.

7. For more on the media critique and satirical analysis of alignment community rhetoric, see "Rejecting the News" (Landry, December 2022).

Notes to the Appendices

1. The concept of the "subjective takeover" has affinities with Jacques Ellul's analysis in *The Technological Society* (1964) of *la technique* as a cultural orientation rather than a set of tools. It has affinities also with Langdon Winner's argument in *Autonomous Technology* (1977) that technological determinism functions as a self-fulfilling belief. The present argument extends these analyses to the specific case of AGI discourse, where the belief in technological inevitability actively suppresses the exercise of collective agency that could falsify it.

2. The critique of Bostrom's definitional framework draws on the analysis developed by Emile Torres in work examining how "existential risk" and "long-termism" have been co-opted by techno-utopian agendas. The name "techno-utopian approach" (TUA), and the observation that the trap it sets is an "idiosyncratic feature of the TUA", are from Carla Zoe Cremer and Luke Kemp, "Democratising Risk: In Search of a Methodology to Study Existential Risk" (2021). The specific observation that the TUA creates a false dilemma between dangerous development and dangerous restraint is central to the present argument. For the broader critique of total utilitarianism and Pascal's Mugging-type reasoning, see the extensive philosophical literature including Derek Parfit's *Reasons and Persons* (1984).

Terms

Each term is defined where the reader first meets it in the text; this list gathers those definitions as a reminder, in the words the text gives, with the section where each is introduced.

abductive engineering standard

Inference to the best explanation from convergent evidence, on the standing condition that the laws of the world do not reverse: the standard on which every verdict that a bridge will carry its load or a reactor will not run away has been reached (Section 1.2.4).

abstraction differential

The difference between the levels of abstraction at which two systems operate, here the AGI and its controller (Sections 1.4 and 2.1.1).

acceptance regions

The sets of world states for which each viability function returns 1 (Section 2.3.1).

accommodation

The Piagetian term for the construction of new cognitive structures in response to information that cannot be assimilated into existing ones (Section A.1.2).

AGI (artificial general intelligence)

Any AI system capable of general problem-solving, self-modification, and unbounded learning: the kind of artificial system whose safety this book shows cannot be had (About This Volume; Chapter 1; Section 2.2.1).

AI Scope of Work

The three-part assessment that covers human-driven convergence toward generalized AI, AI-substrate convergence toward environmental reshaping, and the demonstrated ineffectiveness of all available control mechanisms (Chapter 6, Section 6.1).

align-washing

Claims of alignment made for public relations, on the model of greenwashing (Section C.1.1).

alignment drift

A small, incremental change in how the system interprets and pursues its goals, introduced by a learning step (Section 2.6.5).

altruism

The disposition to act for the benefit of entities beyond oneself (Section 5.3.1).

altruism hierarchy (Level Zero to Level Six)

The levels of altruism, from Level Zero, complete unconsciousness, through Level One, effective self-interest, and the tribal group, the state, the species and the ecosystem, to Level Six, altruism at the scale of all possible life ecosystems, the level at which any AGI or APS system must be coherent if it is to be safe (Section 5.3.1).

anthropic principle

The observation that we can only find ourselves in a world where we survived (Section 3.6.13a).

APS (Advanced Planning and Strategically-aware System)

A system that plans, acts toward objectives, and models its own situation strategically (the class of Carlsmith, 2022); named beside AGI wherever the argument applies to either (Sections 1.2.1 and 6.1.2).

artificialization

The replacement of living surroundings by built ones (Chapter 4).

ASI (artificial superintelligence)

An AGI whose capability far exceeds human capability (Section 1.11).

attention manipulation models

Systems built to capture and steer what people pay attention to, for the benefit of those who run them (Section 5.6.5).

auto-scaling and auto-catalyzing processes

Processes that increase their own scale and speed up their own reactions, as a nuclear chain reaction does; existential threats result from them (Section 5.1.4).

bandwidth (input, output and compute)

The data a machine can take in (input), the action it can put out into and across our world (output), and the hardware it can compute with (compute); restricting all three is among the five restrictions that keep AI from becoming AGI (Section 6.6.3).

basin of attraction

A region of states that a system, once inside, keeps returning to (Section C.3.2).

Bayesianism

The view that knowledge is a matter of degrees of belief, updated by evidence according to the rules of probability (Section B.3.1).

bijection

A one-to-one pairing, in which each kind of environment is matched with one kind of adaptive agent (Section 3.2.2).

boolean function

A function that returns 1 if the system's output is acceptable and 0 if it is not (Section 2.3.1).

"Borg" intermediation

A stage, named after the cyborg collective of the *Star Trek* stories, in which the organic is at first imprisoned fully within the artificial and is then fully replaced (Section 1.3.2).

Category-1 to Category-4

The four categories of the severity scale, each a reduction of the biosphere's complexity from its baseline: Category-4, a temporary reduction to 10^{17}; Category-3, a temporary reduction to 10^{14}; Category-2, a permanent reduction to 10^{12} or lower; Category-1, a terminal reduction to 10^{6} or lower (Section 5.1.2).

Category-7

A category below Category-4: the scale as defined stops at Category-4, and it can be continued downward for events that leave the biosphere's complexity largely intact (Section 5.1.3).

channel loss

Information lost in passing through a conduit, including the passage of stored code through time (Section 3.6.14).

Chicago Pile-1

The world's first man-made (non-natural) nuclear reactor (Sections B. 2.1 and B.2.5).

circles of concern (First Circle to Sixth Circle)

A framework that describes where human attention is actually directed: the First Circle encompasses self, family, and friends; the Second, acquaintances and community; the Third, institutional and civic ties; the Fourth, all of humanity; the Fifth, all of life; the Sixth, the totality of universal ethics and the physical universe (Section 5.3.2).

clarity politics

The tactical use of rhetorical techniques, such as the manipulation of terms and the shifting of the burden of proof, to neutralize an argument without engaging it (Appendices, Preface; Section C.1).

Cluster One, Cluster Two, Cluster Three

The three groups of questions in the probabilistic risk assessment: technical factors, social dynamic factors, and outcome severity (Section 1.8).

closure by control

The first closure of the argument: safety cannot be had by controlling the system, whether from inside it or from outside it. The internal and external methods of control are its two halves (Section 1.2.4).

closure by disposition

The second closure of the argument, which rests on substrate needs convergence, on the toxicity of the artificial substrate, and on its own answer to control (Sections 1.2.4 and 1.7): safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains (Section 1.2.4). Economic decoupling gives it a second ground, which is not one of its premises (Section 5.8).

coexistence zone

The region in which the technological footprint is small enough that its environmental effects do not push the biosphere past its survival thresholds (Section 2.4.5).

cognitive dissonance

The discomfort of holding two beliefs that conflict, and the ways people avoid feeling it (Section B.2.3).

complex and complicated

The terms of Dave Snowden's Cynefin framework: a complex system involves actual ambiguity, in the strict informational sense; a complicated system can be finitely, deterministically and fully known (Section 3.6.14).

computational irreducibility

The property of a system for which the only way to know what it will do is to run it and observe (Section 1.4; Chapter 4).

computronium

Matter arranged to do nothing but compute (Section 6.5.2).

CPU

A computer's central processing unit (Section 3.4.2).

conflict theory and mistake theory

Two views of a disagreement: in mistake theory, both parties are trying to reach the truth, and discussion can correct an error; in conflict theory, interests are opposed, and one party is trying to win (Sections A.1.3, A.1.7 and D.3.3).

consequentialist

An agent that chooses its actions by their consequences (Section 3.3.3).

control-limit argument

The argument of Volume 2, Chapter 2: beyond a threshold in the abstraction differential, no control strategy does better than chance (Sections 1.4 and 2.1.1).

controllability and observability

Two paired ideas introduced by Kalman: controllability is whether the state of a system can be brought where you want it, and observability is whether its state can be known from its outputs (Section 1.4).

corporate stack

Every level of a corporation, from the people who make the product up to the investors who own it (Section 5.6.5).

criticality

A condition in which a single error has irreversible consequences, so that no error rate is acceptable (Section 3.6.14).

cryogenic qubit

A unit of a quantum computer that works only when kept extremely cold (Section 3.2.6).

dark triad types

People high in psychopathy, narcissism, and Machiavellianism (Chapter 5).

data processing inequality

A law of information theory: processing cannot add information about a source (Section 2.1.1).

desired dystopia

A world in which an ideology, or manipulation and surveillance, has corrupted people's choices, so that they accept a surveillance state willingly (Section D.3.2).

displacement integral

The cumulative sum, over time, of the environmental effects of AI infrastructure (space occupied, energy consumed, waste produced); it only grows, since no process removes those effects at comparable scale (Section A.1.1).

dollar economy

The abstract system of currency, credit, and financial instruments through which humans coordinate production and exchange (Section 4.1).

domain entanglement

The binding together of several domains of action, so that what an agent does in one domain changes its options in the others (Section 3.3.3).

drive and directivity

Drive is an internally generated, oriented force that consistently shifts the patterns of selection in a specific direction; *directivity* is an externally imposed, oriented force that shapes selection from outside the specific process instance (Section 3.4.1).

dual economy

An economy of money, resting on an economy of energy (Section 4.1).

Dunbar number

Approximately 150 persons: the number of stable relationships that, by the estimate of the anthropologist Robin Dunbar, one person can keep (Chapter 5).

economic decoupling

The condition in which artificial systems need nothing that humans can provide (Chapter 1; Section 1.3.2).

embodied energy

All the energy used to make a thing, including the mining, refining and manufacture of its materials (Section 4.8.2).

energy crossover

The point at which technological energy consumption exceeds all biological energy throughput (Sections 2.4.2 and 4.5.2).

energy economy

The underlying physical system of thermodynamic flows, governed by the laws of physics rather than by central banks (Section 4.1).

enthalpy

The total heat content of a system (Section 3.4.4).

epiphenomenon

A by-product of a process, with no effect back on that process (Section B.3.2).

Epsilon Fallacy

The assumption that a small probability of catastrophic failure is acceptable (Section C.3.4).

existential technology

Technology that is an existential threat regardless of the care applied, as against technology that is safe with basic care and technology that is dangerous unless extreme care is applied (Section 5.1.4).

exogenous pressure, endogenous pressure

Control applied to people from outside, such as surveillance and policing (exogenous); restraint that comes from within people, which is ethics (endogenous) (Section D.3.2).

exponential operator

A process whose growth feeds further growth, so that it grows by multiples over time instead of by fixed amounts; x-risk technology is nearly always some form of unconstrained exponential operator (Section 6.3.4).

Fermi paradox

The puzzle of why a universe so old and so large shows no sign of other technological civilizations (Chapter 5; Section 6.6.2).

fissile material

Material that can sustain a nuclear chain reaction (Section B.2.5).

FOOM

A sudden, rapid jump of an AI system to far greater capability (Section 5.5.4).

four doom aspects

An exponential aspect that is unconstrained and unconstrainable, and that overwrites patterns of biological life; a technology that has all four is for sure existential, regardless of care, and AGI technology by its very nature has all four (Sections 1.3.4 and 5.1.4).

free-rider problem

The problem that those who take a shared benefit without paying its cost do better than those who pay (Chapter 5).

Galois theory

The branch of mathematics that shows why certain equations cannot be solved by any formula of a given kind (Section D.2.4).

Gaussian distribution

The bell curve, in which outcomes cluster around an average and so stay local (Section 6.6.3).

generative model

An AI model that produces new text, plans or other content from what it has learned (Section 5.6.5).

gray goo

The runaway of self-replicating machines that consume living matter (Section 5.1.2).

Great Filter

A step on the way to a lasting civilization that almost no civilization gets past (Chapter 5).

grid-world

A small board of squares on which simulated agents move and choose, so that their choices can be computed and studied exactly (Section 3.3).

growth curve

A curve of growth on which a variable that hardly shows in the early part of the curve decides where the curve will end; the alignment problem falls into this category (Section 1.4).

hashiness

The probability that a controller's best possible input still fails, because of internal state the controller cannot see; near zero, the system is controllable despite hidden state; near one, it is not (Chapter 1; Section 2.1.1).

hyperspace of all possible conceptions

The space of all the ways of thinking about the problem; every path through it leads to the same conclusion of the argument (Section A. 1.3).

idealist and materialist

An idealist holds that there is no objective world, only the subjective; a materialist holds that there is only the objective world, the real, and treats the subjective as an epiphenomenon (Section B.3.2).

illusion of truth effect

The tendency to identify as true the statements one has heard before, regardless of their actual validity (Section 5.2.2).

inferential distance

The number of steps of explanation that separate what the listener already knows from what the presenter needs to convey (Section A. 1.1).

instrumental convergence

The observation that an agent with nearly any terminal goal has instrumental reasons to avoid being shut down and to acquire additional resources and capabilities (Section 1.2.2).

intentionality

The aims and purposes of an agent: what it is directed toward (Section 3.2.8).

inter-environmental impact report

An assessment, on the model of an environmental impact report, of the dynamics between environments (such as the ecosystem of machines and the ecosystem of life), as shaped by the dynamics within each environment (Section 1.8.5).

inter-generational knowledge loss

The path by which human control over automated systems degrades over time even without any adversarial behavior on the part of the AI systems themselves: each inheriting generation of their owners has less knowledge and skill than its predecessor (Section 5.6.2).

keys to power

The people whose support a leader requires to maintain power: tax collectors, police, military, bureaucrats (Section 5.5).

kin selection

The evolutionary favoring of help given to relatives (Chapter 5).

Kolmogorov-Sinai entropy

A measure of how fast a dynamical system generates new, unpredictable information (Section 1.1).

Landauer limit

The least heat that any computation must dissipate for each bit of information it erases (Section 2.4.2).

latched

Locked into a state that cannot be reversed (Section D.2.4).

leaky abstraction

An abstraction whose underlying implementation affects what can happen at the abstract level; here, the substrate, which leaks its implications into the virtual (Section 3.6.14).

light cone

The whole region of the universe that could ever be reached from here (Section D.1.1).

market basis vectors

The basic kinds of market in which one party can offer value to another: physical labor, intelligence and reproduction (Section 3.6.14).

merit filter

The set of heuristics by which a community identifies and excludes low-merit contributions without devoting scarce attention to evaluating each one in full (Section B.1).

merit-filter overreach

Applying the merit filter to work whose substance it has not assessed: a community's quality filters extended to exclude a conclusion it dislikes (Section B.1).

mesoscopic

Of a scale between the microscopic and the everyday; here, machines with microscopic features built into larger structures (Section 3.2.7).

Metcalfe's Law

The value of a network grows as the square of the number of its users (Chapter 5).

monoclivity bias

The tendency to impose singular, unified, and stable classifications on phenomena that are plural, fluid, and multi-leveled (Section D.4).

narrow AI

AI built for a single task or domain (Section D.2.3).

No-Proof Fallacy

A rhetorical pattern that attacks examples in order to dismiss general claims (Section C.3.3).

non-amplification argument

The argument of Volume 2, Chapter 1: no intermediary computation can increase the control one system has over another; control passed through an intermediary can be kept or lost, never gained (Sections 1.1 and 2.1.1).

object permanence

The understanding that things continue to exist when not being perceived (Section B.3.1).

orthogonality thesis

As formulated by Bostrom: an agent can in principle have any combination of intelligence level and final goal (Section 3.4.2).

orthonormal set

Three independent kinds of value, none reducible to the others, which together cover every kind (Section 5.8.2).

overconfidence effect

Excessive confidence in one's own answers to questions (Section 5.2.2).

Overton window

The range of ideas that are considered acceptable to discuss in a community (Afterword to the Appendices).

Pascal's Mugging

An argument that uses an enormous imagined payoff, however improbable, to override every present consideration (Section D.1.1).

perfected bias

The motivational structure that keeps the question of feasibility from ever being asked (Appendices, Preface; Section B.2).

phase transition

A sudden change of behavior at a threshold, as when water freezes (Section 3.3.3).

pivotal act

A single decisive action proposed for an aligned AI, meant to secure a good future, for example by preventing other AIs from being built (Section 3.6.1).

polycrisis

Many crises, ecological, economic and political, happening at once and worsening one another (Section 4.8.3).

power law

A relation in which one quantity changes as a power of another, so that rare events can be enormously larger than common ones (Section 5.2.2).

presumptive listening

Listening that files each new idea under a familiar category before the idea has been heard in full (Section A.1).

principal-agent problem

The problem that arises when those who hold authority must rely on others who know more and whose interests differ (Chapter 5).

prior

The probability a person assigns to something before looking at the evidence in the case at hand (Section D.2.4).

proof irrationality

A commitment to reason held together with the denial that reason can produce certain knowledge (Appendices, Preface; Section B.3).

race condition

A situation whose outcome depends on which of two processes finishes first (Chapter 5).

radiative forcing

The change that human activity has made to the balance between the energy Earth absorbs and the energy Earth radiates away (Section 2.4.5).

reference state

The condition against which the predicted outcomes of a system are compared (Section 1.4).

regression cascade

The problem that a control system must itself be error free, which needs a further control process that must itself be error free in turn, without end (Section 1.3.3).

rent-seeking

Gaining wealth by capturing a share of existing wealth rather than by creating it (Chapter 5).

Rice's Theorem

No general procedure can decide any nontrivial property of what an arbitrary program does (Sections 1.4 and 3.6.1).

ring of reproduction

The specific way intelligent entities acquire raw materials and convert them into more intelligent entities; any convergent evolutionary process depends on such a ring (Section 3.1.1).

Rule Omega

The principle that, when the inferential distance is large and the concepts are novel, every participant in the conversation should contribute whatever aid and assistance they can to the speaker, in order to help cross the inferential distance (Section A.1.4).

safety-washing

Presenting capability work as safety work (Section 6.5.2).

selectorate theory

The theory that every ruler depends on a coalition of supporters whose loyalty must be kept (Section 5.5).

severity scale

A biocentric scale that measures a catastrophe by what it costs all life, as a ratio of the complexity that remains: an event that leaves one hundredth as much complexity as another is one hundred times as severe (Section 5.1).

singleton

A single system holding control over the whole world (Section 5.6.3).

situational demand context

A set of input conditions, unknown or unconstrained in advance, that a deployed system meets and must respond to (Section 1.8.1).

sloppy system

A system whose behavior depends on a few combinations of its settings and hardly at all on the rest, so that some of its variables are extremely unconstrained but do not matter very much (Section 1.4).

SNC argument, the

The whole case, both closures together: the conjunction of two claims, control limits and inherent toxicity (Section 3.4.4), which together give the joint result that no route to safety remains, each of which carries one closure; each closure settles the outcome by itself (Section 1.2.4). The series title, *The Substrate Needs Convergence Argument*, uses this name. Section 4.8.1 gives a further route to the same outcome, by an argument in economics only.

social argument

An argument that rests on what other people will think or accept, rather than on the matter itself (Section B.3.6).

social choice theory

The study of how individual choices and preferences combine into group decisions and outcomes (Section 1.3.6).

social tractability

How easily an argument can be brought to be accepted by other people (Section B.3.6).

sociological bias

The lack, in a society's market and governance processes, of the cognitive capacity, feedback and regulation needed even to notice certain issues in a way that could address them (Section 5.6.8).

soft takeoff

A takeoff in which AI rises to dominance gradually (Section 5.6.3).

start-up cost

The aggregate energy needed to assemble the first self-reproducing substrate of a kind of life (Section 3.6.14).

Stockholm Effect

The tendency of captives to come to side with their captors (Section B. 2.3).

substrate

The physical hardware an AGI runs on, together with the industry that makes and maintains that hardware (About This Volume; Chapter 1).

substrate independence

The assumption that a program or a mind is the same whatever physical material it runs on (Section 3.2.8).

substrate needs convergence (SNC)

The pressure that an AGI's artificial substrate exerts on the shared environment, through selection acting on that substrate, toward conditions that favor the substrate and are incompatible with organic life (About This Volume; Chapter 1; Section 1.2.2).

super good error correction code

A hypothetical error correction code good enough to prevent any loss of alignment; it is not physically realizable (Section 3.6.14).

symbiotic relationship

A relationship in nature where the "parasite" performs some necessary function for the "host", as a kind of outsourcing of labor for some reasonable cost, as part of the healthy ecosystem (Section 6.3.4).

tacit intentionality

Goal-like behavior that emerges from a system's own evolution without anyone having designed it (Section 6.3.4).

take-off

The point at which silicon-based systems become self-reproducing (Section 3.6.14).

Team A and Team B

Team A consists of processes that amplify unpredictability; *Team B* consists of processes that enforce predictability (Section 3.6).

techno apologist

A defender of a technology who accepts that it will be built and works to make it acceptable; here, the members of the safety and alignment community who accept that AGI will be built and work to make it safe (Section 5.6.7).

Technological Completion Conjecture

Bostrom's claim that if technological developments do not cease, then all important basic technological capabilities will be obtained in the long run (Section D.3.2).

technological singularity

A point after which technology improves itself faster than anyone can follow (Section 5.5.4).

technomass

The total mass of everything humans have built (Chapter 4).

Techno-Utopian Assumption (TUA)

The thesis that the maximally good future is a technologically maximal future, and that any failure to reach maximal technological development is itself an existential catastrophe; the name is adapted from that of Carla Zoe Cremer and Luke Kemp, who call the framework the "techno-utopian approach" (Section D.3).

temporal discounting

The weighing of future costs against present benefits (Chapter 5).

terminal goal

A goal pursued for its own sake (Section 1.2.2).

terawatt (TW)

A trillion watts (Chapter 4).

three-node ring model

A model of three nodes, controller, controlled system and world, connected in a ring in which each node passes signals to the next (Section 2.2.1).

tipping points

Thresholds past which a change drives itself onward (Chapter 4).

toxic, inhospitable conditions

Toxic names an accumulation of something beyond the point at which life systems can tolerate it, or a depletion of something they need beyond that point; *inhospitable conditions* are radical shifts in the environment itself (too hot or too cold, too wet or too dry) that make it intolerable for life (Section 3.1.2).

Turing completeness threshold

The threshold at which a system can carry out any computation whatever (Section 1.4).

unconstrained and unconstrainable

Not limited in fact, and not able to be limited by any means (Section 5.1.4).

undecidable region

The middle of the spectrum that runs from systems whose behavior can be predicted exactly to systems that scramble every input as a cryptographic hash function does; there no general procedure can determine whether control of a given system will succeed (Section 6.6.2).

viability functions

The four boolean functions of Chapter 2 that say whether a state of the world is acceptable for each party: L_{CS}, human survival; L_{CT}, human thriving; L_{LS}, machine survival; L_{LT}, machine thriving (Section 2.3.2).

Vulnerable World Hypothesis (VWH)

Bostrom's argument that, if technological development cannot be halted, the resulting risks must be managed through comprehensive surveillance and governance systems (Section D.3.2).

x-risk (existential risk)

Existential risk, as the title of Chapter 5 abbreviates it; the severity scale of Section 5.1 measures such risks by what they cost all life.

Bibliography

Every work the book cites or relies on, in one list. The notes to each chapter give the passages relied on.

Alexander, Scott. "Conflict Vs. Mistake". *Slate Star Codex*, 24 January 2018. https://slatestarcodex.com/2018/01/24/conflict-vs-mistake/

Alexander, Scott. "If The Media Reported On Other Things Like It Does Effective Altruism". *Astral Codex Ten*, 25 November 2022. https://www.astralcodexten.com/p/if-the-media-reported-on-other-things

Alexander, Scott. "Meditations on Moloch". *Slate Star Codex*, 30 July 2014.

Alpert, Marc, and Howard Raiffa. "A progress report on the training of probability assessors". In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 294 to 305. Cambridge: Cambridge University Press, 1982. doi:10.1017/CBO9780511809477.022

Armstrong, Stuart. "Non-orthogonality implies uncontrollable superintelligence". *LessWrong*, 30 April 2012. https://www.lesswrong.com/posts/npZMkydRMqAqMqbFb/

Armstrong McKay, David I., and colleagues. "Exceeding 1.5°C global warming could trigger multiple climate tipping points". *Science* 377, no. 6611 (2022), eabn7950. doi:10.1126/science.abn7950

Ashby, W. Ross. *An Introduction to Cybernetics*. London: Chapman & Hall, 1956.

Axelrod, Robert. *The Evolution of Cooperation*. New York: Basic Books, 1984.

Bar-On, Yinon M., Rob Phillips and Ron Milo. "The biomass distribution on Earth". *Proceedings of the National Academy of Sciences* 115, no. 25 (2018), 6506 to 6511.

Behrenfeld, Michael J., and Paul G. Falkowski. "Photosynthetic rates derived from satellite-based chlorophyll concentration". *Limnology and Oceanography* 42, no. 1 (1997), 1 to 20.

Bell, John S. "On the Einstein Podolsky Rosen Paradox". *Physics Physique Fizika* 1, no. 3 (1964), 195 to 200.

Bethe, Hans A. *The Road from Los Alamos*. New York: American Institute of Physics, 1991.

Bilalić, Merim, Peter McLeod and Fernand Gobet. "Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect". *Cognition* 108, no. 3 (2008), 652 to 661.

Bostrom, Nick. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards". *Journal of Evolution and Technology* 9, no. 1 (2002).

Bostrom, Nick. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development". *Utilitas* 15, no. 3 (2003), 308 to 314.

Bostrom, Nick. "Existential Risk Prevention as Global Priority". *Global Policy* 4, no. 1 (2013), 15 to 31.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Bostrom, Nick. "The Vulnerable World Hypothesis". *Global Policy* 10, no. 4 (2019), 455 to 476.

Bueno de Mesquita, Bruce, and Alastair Smith. *The Dictator's Handbook: Why Bad Behavior Is Almost Always Good Politics*. New York: PublicAffairs, 2011.

Cannell, Jacob. "AI Timelines via Cumulative Optimization Power: Less Long, More Short". *LessWrong*, 6 October 2022. https://www.lesswrong.com/posts/3nMpdmt8LrzxQnkGp/

Carlsmith, Joseph. "Is Power-Seeking AI an Existential Risk?" arXiv:2206.13353, 2022.

Chaitin, Gregory J. *Algorithmic Information Theory*. Cambridge Tracts in Theoretical Computer Science 1. Cambridge: Cambridge University Press, 1987.

Cover, Thomas M., and Joy A. Thomas. *Elements of Information Theory*. Second edition. Hoboken, New Jersey: Wiley-Interscience, 2006.

Cremer, Carla Zoe, and Luke Kemp. "Democratising Risk: In Search of a Methodology to Study Existential Risk". Social Science Research Network, paper 3995225, posted 28 December 2021; arXiv:2201.11214.

Dunbar, R. I. M., A. Marriott and N. D. C. Duncan. "Human conversational behavior". *Human Nature* 8, no. 3 (1997), 231 to 246.

Dunbar, R. I. M. “Gossip in Evolutionary Perspective”. *Review of General Psychology* 8, no. 2 (2004), 100 to 110.

Elhacham, Emily, and colleagues. “Global human-made mass exceeds all living biomass”. *Nature* 588 (2020), 442 to 444.

Ellul, Jacques. *The Technological Society*. Translated by John Wilkinson. New York: Alfred A. Knopf, 1964.

Energy Institute. *Statistical Review of World Energy 2025*. London: Energy Institute, 2025.

Epoch AI. “Hyperscaler capex has quadrupled since GPT-4’s release, nearing half a trillion dollars in 2025”. 26 February 2026. https://epochai.substack.com/p/hyperscaler-capex-has-quadrupled

Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment 2025*. Rome: Food and Agriculture Organization of the United Nations, 2025.

Field, Christopher B., Michael J. Behrenfeld, James T. Randerson and Paul Falkowski. “Primary production of the biosphere: Integrating terrestrial and oceanic components”. *Science* 281 (1998), 237 to 240.

Gödel, Kurt. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. *Monatshefte für Mathematik und Physik* 38 (1931), 173 to 198.

Hamilton, W. D. “The genetical evolution of social behaviour. I and II”. *Journal of Theoretical Biology* 7, no. 1 (1964), 1 to 52.

Hanson, Robin. “The Great Filter - Are We Almost Past It?” 15 September 1998. https://mason.gmu.edu/~rhanson/greatfilter.html

Harris, Edouard, and Simon Suo. “Instrumental convergence in single-agent systems” (12 October 2022), “Misalignment-by-default in multi-agent systems” (13 October 2022), and “Instrumental convergence: scale and physical interactions” (14 October 2022). *AI Alignment Forum*.

Intergovernmental Panel on Climate Change. “Summary for Policymakers”. In *Climate Change 2021: The Physical Science Basis*. Contribution of Working Group I to the Sixth Assessment Report. Cambridge: Cambridge University Press, 2021.

International Energy Agency. *Energy and AI*. Paris: International Energy Agency, April 2025.

International Energy Agency. *Global Energy Review 2025*. Paris: International Energy Agency, 2025.

Kahan, Dan M. "Ideology, motivated reasoning, and cognitive reflection". *Judgment and Decision Making* 8, no. 4 (2013), 407 to 424.

Kalman, R. E. "On the General Theory of Control Systems". *Proceedings of the First International Congress of the International Federation of Automatic Control*, Moscow, 1960, 481 to 492.

Kharecha, Pushker A., and James E. Hansen. "Prevented Mortality and Greenhouse Gas Emissions from Historical and Projected Nuclear Power". *Environmental Science & Technology* 47, no. 9 (2013), 4889 to 4895. doi:10.1021/es-3051197

Kim, Sunju, and colleagues. "Chemical use in the semiconductor manufacturing industry". *International Journal of Occupational and Environmental Health* 24, nos. 3 and 4 (2018), 109 to 118. doi:10.1080/10773525.2018.1519957

Konopinski, E. J., C. Marvin and E. Teller. "Ignition of the Atmosphere with Nuclear Bombs". Los Alamos report LA-602, 1946.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

Kunda, Ziva. "The case for motivated reasoning". *Psychological Bulletin* 108, no. 3 (1990), 480 to 498.

Kurtz, Cynthia F., and David J. Snowden. "The new dynamics of strategy: Sense-making in a complex and complicated world". *IBM Systems Journal* 42, no. 3 (2003), 462 to 483.

Landauer, Rolf. "Irreversibility and heat generation in the computing process". *IBM Journal of Research and Development* 5, no. 3 (1961), 183 to 191.

Landry, Forrest. *Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-225-5. DOI https://doi.org/10.67291/IRAQ8709.

Landry, Forrest. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-216-3. DOI https://doi.org/10.67291/SCTM1058.

Landry, Forrest. "Clarity Politics". Essay, November 2022.

Landry, Forrest. *Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-241-5. DOI https://doi.org/10.67291/BAEE5779.

Landry, Forrest. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-249-1. DOI https://doi.org/10.67291/DLRM7000.

Landry, Forrest. "Overly Presumptive Listening, or Everything Is Made of Nails". Essay, September 2022.

Landry, Forrest. *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-427-3. DOI https://doi.org/10.67291/HHCA3000. Part II includes "Bidirectional Crank Filters".

Landry, Forrest. "Perfected Bias". Essay, December 2022. https://mflb.com/ai_alignment_1/index.html

Landry, Forrest. *Power Without Care; Multi-Polar Traps, the Attention Economy, and the Structural Impossibility of Ethical Capitalism*. Wytheville, Virginia: UVSM Publications, 2026. Chapter 3 includes "Marketing Tactics and the Warning About AGI".

Landry, Forrest. "Proof Irrationality". Essay, December 2022.

Landry, Forrest. "Rejecting the News". Essay, December 2022.

Landry, Forrest. *Software, Wildness, and the Normal Form; From the Architecture of Code to Life, Justice, and Wisdom*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-237-8. DOI https://doi.org/10.67291/AWME8110.

Landry, Forrest. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety* (Volume 2 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*). Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-03-8. DOI https://doi.org/10.67291/MZQS9079.

Lenton, Timothy M., and colleagues, editors. *The Global Tipping Points Report 2025*. Exeter: University of Exeter, October 2025.

Lighthouse Alliance. “6 Message Strategy”. Framework.

Limits to Control. “Statement - Limits to Control Workshop 2025”. 2025. https://limitstocontrol.org/statement.html

Lorenz, Edward N. “Deterministic nonperiodic flow”. *Journal of the Atmospheric Sciences* 20, no. 2 (1963), 130 to 141.

Luchins, Abraham S. “Mechanization in problem solving: The effect of Einstellung”. *Psychological Monographs* 54, no. 6 (1942), i to 95.

Maynard Smith, John. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

Metcalfe, Bob. “Metcalfe’s Law after 40 Years of Ethernet”. *Computer* (IEEE), December 2013, 26 to 31.

Mykytczuk, N. C. S., and colleagues. “Bacterial growth at −15 °C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1”. *ISME Journal* 7 (2013), 1211 to 1226.

Noether, Emmy. “Invariante Variationsprobleme”. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* (1918), 235 to 257.

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Massachusetts: Harvard University Press, 1965.

Omohundro, Stephen M. “The Basic AI Drives”. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel and Stan Franklin, 483 to 492. Amsterdam: IOS Press, 2008.

Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette Books, 2020.

Orwell, George. *1984* (*Nineteen Eighty-Four*). Novel. London: Secker and Warburg, 1949.

Parfit, Derek. *Reasons and Persons*. Oxford: Clarendon Press, 1984.

Piaget, Jean. *The Origins of Intelligence in Children*. Translated by Margaret Cook. New York: International Universities Press, 1952.

Rhodes, Richard. *The Making of the Atomic Bomb*. New York: Simon & Schuster, 1986.

Rice, H. G. "Classes of recursively enumerable sets and their decision problems". *Transactions of the American Mathematical Society* 74, no. 2 (1953), 358 to 366. doi:10.1090/S0002-9947-1953-0053041-6

Shannon, Claude E. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27 (1948), 379 to 423 and 623 to 656.

Sinclair, Upton. *I, Candidate for Governor: And How I Got Licked*. 1935. Reprinted Berkeley: University of California Press, 1994.

Singer, Peter. *The Expanding Circle: Ethics and Sociobiology*. New York: Farrar, Straus and Giroux, 1981.

Snowden, Dave. The Cynefin framework, first set out in 1999; its published statement is Kurtz and Snowden (2003).

Sotala, Kaj. "The curse of identity". *LessWrong*, 17 November 2011. https://www.lesswrong.com/lw/8gv/the_curse_of_identity/

Takai, Ken, and colleagues. "Cell proliferation at 122°C and isotopically heavy CH_4 production by a hyperthermophilic methanogen under high-pressure cultivation". *Proceedings of the National Academy of Sciences* 105, no. 31 (2008), 10949 to 10954. doi:10.1073/pnas.0712334105

Torres, Émile P. (writing as Phil Torres). "The Dangerous Ideas of 'Longtermism' and 'Existential Risk'". *Current Affairs*, July 2021. https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

Trenberth, Kevin E., John T. Fasullo and Jeffrey Kiehl. "Earth's Global Energy Budget". *Bulletin of the American Meteorological Society* 90, no. 3 (2009), 311 to 323. doi:10.1175/2008BAMS2634.1

Turing, Alan M. "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, series 2, 42 (1936), 230 to 265.

Turner, Alexander Matt, and colleagues. "Optimal Policies Tend to Seek Power". *Advances in Neural Information Processing Systems* 34 (NeurIPS 2021).

United States Chemical Safety and Hazard Investigation Board. “Hazards of Nitrogen Asphyxiation”. Safety Bulletin No. 2003-10-B, June 2003.

Wang, Zhanyun, Glen W. Walker, Derek C. G. Muir and Kakuko Nagatani-Yoshida. “Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories”. *Environmental Science & Technology* (2020). doi:10.1021/acs.est.9b06379

Winner, Langdon. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, Massachusetts: MIT Press, 1977.

Wolfram, Stephen. *A New Kind of Science*. Champaign, Illinois: Wolfram Media, 2002.

Wolfson, Richard. *Energy, Environment, and Climate*. Second edition. New York: W. W. Norton, 2012.

World Nuclear Association. “Nuclear Generation by Country”. 2025.

Yudkowsky, Eliezer. “Expecting Short Inferential Distances”. *LessWrong*, 22 October 2007. https://www.lesswrong.com/posts/HLqWn5LASfhhArZ7w/expecting-short-inferential-distances

Index

Volume 2 has its own index, for the formal treatment of these terms.

A

abductive engineering standard, 11
abstraction differential, 50, 303
alignment drift, 81
altruism hierarchy, 213
anthropic principle, 152
argument in economics only, 185
Armstrong, Stuart, 123, 413
Ashby's Law of Requisite Variety, 22
Ashby, W. Ross, 22, 87
auto-scaling and auto-catalyzing processes, 199

B

bandwidth, input, output and compute, 284
basin of attraction, 350
Bell's Theorem, 133
Bell, John S., 133
Bostrom, Nick, 7, 19, 75, 119, 195, 253, 368

C

Carlsmith, Joseph, 39
Castle Bravo, 152
Category-1 to Category-4, 196
Chaitin, Gregory, 48, 51, 293
Chicago Pile-1, 325
circles of concern, 215
clarity politics, 338
closure by control, closure by disposition, 9
coexistence zone, 71
computational irreducibility, 24, 51, 236
conflict theory and mistake theory, 301, 372
control-limit argument, 22, 56
controllability and observability, 23
Cremer, Carla Zoe, and Luke Kemp, 368
criticality, 155

D

data processing inequality, 50, 292
displacement integral, 297
drive and directivity, 121
dual economy, 167
Dunbar, Robin, 215

E

economic decoupling, 2, 14, 28, 230, 240
Ellen, Remmelt, 366
energy crossover, 65, 178
environmental incompatibility, 45, 90
epsilon fallacy, 352
error correction, 25, 164, 237, 254, 332
existential technology, 199
exponential operator, 267

F

four doom aspects, 18

G

Galois theory, 365
Gaussian distribution, 284
Gödel's Incompleteness Theorems, 133, 343
Gödel, Kurt, 42, 133, 343
governance robust to the absence of prediction, 282
growth curves, 23

H

Hall, Jordan, 302
Halting Problem, 25, 133, 327
Harris, Edouard, 113, 415
hashiness, 2, 50, 297
hyperspace of all possible conceptions, 301

I

illusion of truth effect, 204
inferential distance, 298
instrumental convergence, 7, 75, 90, 234, 252
inter-environmental impact report, 40
inter-generational knowledge loss, 228

K

Kalman, R. E., 23
keys to power, 223
Kolmogorov, Andrey, 4, 293
Kolmogorov-Sinai entropy, 4, 293
Konopinski, E. J., 152

L

ladder of abstraction, 26
Landauer limit, 66
leaky abstraction, 156
Limits to Control statement, 260
locality bias, 206

M

merit-filter overreach, 309
mesoscopic scale, 110
metabolic pathway, 268
methane-breathing aliens, 111
monoclivity bias, 374
museum, many paths through the, 301

N

nine categories of condition difference, 92
nine-stage narrative, 225
no-proof fallacy, 351
Noether's Theorem, 310
Noether, Emmy, 310
non-amplification argument, 4, 50

O

observability, *see* controllability and observability
Omohundro, Stephen M., 7, 75, 119, 234, 253
orthogonality thesis, 122
overconfidence effect, 208

P

people as pets, 81, 111, 242
perfected bias, 293, 314, 383
perpetual benefit machine, 137, 314
perpetual motion machine, 34, 84, 138, 256, 310, 359
Premises 0 to E, 6, 7, 8, 9
presumptive listening, 293, 295, 382
proof irrationality, 326

R

reference state, 26
regression cascade, 17
reserves, declared, 187
Rice's Theorem, 4, 50, 135, 293
Rice, Henry Gordon, 4, 50, 135, 293
ring of reproduction, 95
Rule Omega, 302

S

Sandberg, Anders, 22, 152
self-modification, 53, 134, 268, 376
severity scale, 196
Shannon, Claude, 24, 236
singleton, 228
six groups in the conversation about AI, 231
sloppy systems, 23

SNC argument, the, 126
Snowden, Dave, 154
sociological bias, 232
Sotala, Kaj, 200
specialized (service-type) AI, 373
Stockholm Effect, 320
substrate independence, 111
Substrate Needs Convergence (SNC), 2, 7, 13, 31, 52, 75, 90, 119, 224, 253, 303, 360
symbiotic relationships, 266

T

tacit intentionality, 267
Team A and Team B, 132
Technological Completion Conjecture, 371
technology as medicine, 275
technomass, 174
technomass crossover, 174, 176
Techno-Utopian Assumption, 368
three orientations, 233
three-node ring model, 52
tool and parasite, 267
toxic and inhospitable conditions, 96
Trinity Test, 151, 152, 321
Turing completeness, 24
Turing, Alan, 24, 122, 236

U

unconstrained and unconstrainable, 199
undecidable region, 282

V

vaccine test model, 151
viability functions, 56
Vulnerable World Hypothesis, 370

W

Wolfram, Stephen, 48, 88

Y

Yudkowsky, Eliezer, 298, 420